UNIVERSIDADE FEDERAL DE PELOTAS
Centro de Desenvolvimento Tecnológico
Programa de Pós-Graduação em Computação

Tese

Desenvolvimento de Abordagens Baseadas em Redes Neurais Profundas para Detecção e Segmentação de Instância de Lesões Retinianas

Carlos Alexandre Silva dos Santos

Pelotas, 2023

Carlos Alexandre Silva dos Santos

Desenvolvimento de Abordagens Baseadas em Redes Neurais Profundas para Detecção e Segmentação de Instância de Lesões Retinianas

Tese apresentada ao Programa de Pós-Graduação em Computação do Centro de Desenvolvimento Tecnológico da Universidade Federal de Pelotas, como requisito parcial à obtenção do título de Doutor em Ciência da Computação.

Orientador: Prof Dr. Marilton Sanchotene de Aguiar
Coorientador: Prof Dr. Daniel Welfer

Pelotas, 2023

Universidade Federal de Pelotas / Sistema de Bibliotecas
Catalogação da Publicação

S237d Santos, Carlos Alexandre Silva dos

Desenvolvimento de abordagens baseadas em redes neurais profundas para detecção e segmentação de instância de lesões retinianas [recurso eletrônico] / Carlos Alexandre Silva dos Santos ; Marilton Sanchotene de Aguiar, orientador ; Daniel Welfer, coorientador. — Pelotas, 2023.
216 f.

Tese (Doutorado) — Programa de Pós-Graduação em Computação, Centro de Desenvolvimento Tecnológico, Universidade Federal de Pelotas, 2023.

1. Retinopatia diabética. 2. Imagens de fundo. 3. Aprendizado profundo. 4. Detecção de lesões de fundo. 5. Segmentação de instância de lesões de fundo. I. Aguiar, Marilton Sanchotene de, orient. II. Welfer, Daniel, coorient. III. Título.

CDD 005

Elaborada por Simone Godinho Maisonave CRB: 10/1733

Carlos Alexandre Silva dos Santos

Desenvolvimento de Abordagens Baseadas em Redes Neurais Profundas para Detecção e Segmentação de Instância de Lesões Retinianas

Tese aprovada, como requisito parcial, para obtenção do grau de Doutor em Ciência da Computação, Programa de Pós-Graduação em Computação, Centro de Desenvolvimento Tecnológico, Universidade Federal de Pelotas.

Data da Defesa: 15 de dezembro de 2023.

Banca Examinadora:

Prof. Dr. Marilton Sanchotene de Aguiar (orientador)
Doutor em Computação pela Universidade Federal do Rio Grande do Sul.

Prof. Dr. Daniel Welfer (coorientador)
Doutor em Computação pela Universidade Federal do Rio Grande do Sul.

Prof. Dr. Bruno Zatt
Doutor em Microeletrônica pela Universidade Federal do Rio Grande do Sul.

Prof. Dr. Daniel Fernando Tello Gamarra
Doutor em Informática pela Scuola Superiore Sant'Anna di Studi Universitari e Perfezionamento.

Prof. Dr. João Mário Lopes Brezolin
Doutor em Ciência da Computação pela Pontifícia Universidade Católica do Rio Grande do Sul.

Pelo carinho, dedicação e sabedoria dedico esta Tese a meus pais e a todos meus queridos professores que tive desde meu ensino fundamental até a pós-graduação, sem os quais nada disso seria possível.

AGRADECIMENTOS

Primeiro, a Deus por iluminar meu caminho. Sempre senti sua presença, nos bons e maus momentos de minha vida. Nos momentos de dor, dúvida ou desespero estava lá sob alguma forma, dando-me forças para resistir, e sinalizando-me o que fazer.

À minha família que sempre esteve ao meu lado, e que me ensinou todos os princípios fundamentais para que me tornasse um homem digno.

À minha esposa que sempre esteve ao meu lado. Acreditou em mim e nos meus sonhos, trabalhando junto comigo para que estes se tornassem realidade.

A todos meus amigos, que infelizmente não poderei nominar aqui, que sempre torceram por mim, e que na hora que mais precisei estenderam suas mãos para me apoiar. Vocês estão sempre em minhas orações. Saibam que podem contar comigo hoje e sempre.

Aos meus orientadores, professores Marilton Aguiar e Daniel Welfer, pelo apoio e confiança a mim depositada durante a pesquisa.

À médica oftalmologista e professora do Departamento de Medicina Especializada da Universidade Federal de Pelotas, Tatiana Papaléo, pela colaboração durante a pesquisa.

Ao médico oftalmologista José Américo Pascal Proto, da Clínica Proto Oftalmologia da cidade de Pelotas/RS, pela colaboração durante a pesquisa.

Aos professores Marcelo Porto, Bruno Zatt e Ricardo Araujo, do Programa de Pós-Graduação em Computação da Universidade Federal de Pelotas, pela atenção e suporte que sempre me forneceram ao longo do curso.

Aos membros do laboratório de Aprendizado de Máquina, Inteligência Artificial e Ciência de Dados ligado ao Grupo de Pesquisa em Ciência de Dados na Universidade Federal de Pelotas (Datalab/ML–UFPel), pela disponibilização de equipamentos para a realização de parte dos experimentos da pesquisa.

Ao Bruno Belloni do Instituto Federal Sul-Rio-Grandense (IFSul), e ao Marcelo Dias, Alejandro Pereira e Fernando Ollé, da Universidade Federal de Pelotas (UFPel), pela colaboração durante a pesquisa.

Ao André Luiz Almeida Silva do Instituto Tecnológico de Aeronáutica (ITA), pela colaboração durante a pesquisa.

Ao professor Prasanna Porwal do Instituto de Engenharia e Tecnologia Shri Guru Gobind Singhji, Nanded, Índia, pela colaboração durante a pesquisa.

Por fim, aos meus colegas do Instituto Federal Farroupilha, Câmpus Alegrete, pela camaradagem, profissionalismo e compreensão quanto à distribuição de funções a mim atribuídas, em virtude da impossibilidade de concessão de afastamento integral para a realização do doutorado.

Healthcare is About Humans Caring For Humans.
AI is Here to Help Enhance Humanity.
— Fei-Fei Li
EMBC'21

RESUMO

SANTOS, Carlos Alexandre Silva dos. **Desenvolvimento de Abordagens Baseadas em Redes Neurais Profundas para Detecção e Segmentação de Instância de Lesões Retinianas**. Orientador: Marilton Sanchotene de Aguiar. 2023. 216 f. Tese (Doutorado em Ciência da Computação) – Centro de Desenvolvimento Tecnológico, Universidade Federal de Pelotas, Pelotas, 2023.

A Retinopatia Diabética (RD) é uma das principais causas de perda de visão e apresenta em suas fases iniciais lesões de fundo, como microaneurismas, hemorragias e exsudatos duros e algodonosos. Modelos computacionais capazes de detectar essas lesões podem auxiliar no diagnóstico precoce da doença e prevenir a manifestação de formas mais graves de lesões, auxiliando também no processo de triagem e definição da melhor forma de tratamento. Entretanto, a detecção de microlesões por meio de sistemas computacionais é um desafio por inúmeros fatores, como o tamanho e formato destas lesões, a presença de ruído e contraste ruim das imagens, a pequena quantidade de exemplos rotulados nos conjuntos de dados públicos de RD, e a dificuldade de algoritmos de aprendizado profundo em detectar objetos muito pequenos em função da dissipação de gradiente durante o treinamento. Assim, para contornar estes problemas, este trabalho propõe duas novas abordagens baseadas em técnicas de processamento de imagens, aumento de dados, transferência de aprendizado e redes neurais profundas, com o propósito de auxiliar no diagnóstico médico de lesões de fundo. As abordagens propostas foram treinadas, ajustadas e avaliadas usando diferentes conjuntos de dados públicos de Retinopatia Diabética. Para a realização dos experimentos os *datasets* foram particionados em conjunto de treinamento (50%), validação (20%) e teste (30%). Utilizou-se uma etapa de validação para realizar o ajuste fino de hiperparâmetros, e uma etapa de teste para aferir a capacidade de generalização dos modelos. A abordagem para detecção das lesões de fundo alcançou mAP de 0,2630 para o limite de IoU de 0,5 na etapa de validação utilizando o conjunto de dados DDR e otimizador Adam. Já a abordagem para segmentação de instância das lesões de fundo alcançou mAP de 0,2903 para o limite de IoU de 0,5 na etapa de validação utilizando o conjunto de dados DDR e otimizador Adam, sendo, portanto, 10,38% mais preciso que a abordagem proposta para detecção. Os resultados obtidos nos experimentos demonstram que as novas abordagens apresentaram resultados promissores na detecção de lesões de fundo associadas à RD.

Palavras-chave: Retinopatia Diabética. Imagens de Fundo. Aprendizado Profundo. Detecção de Lesões de Fundo. Segmentação de Instância de Lesões de Fundo.

ABSTRACT

SANTOS, Carlos Alexandre Silva dos. **Development of Deep Neural Network-Based Approaches for Detection and Instance Segmentation of Retinal Lesions**. Advisor: Marilton Sanchotene de Aguiar. 2023. 216 f. Thesis (Doctorate in Computer Science) – Center for Technological Development, Federal University of Pelotas, Pelotas, 2023.

Diabetic Retinopathy (DR) is one of the leading causes of vision loss and presents fundus lesions in its initial stages, such as microaneurysms, hemorrhages, hard exudates, and soft exudates. Computational models capable of detecting these lesions can support the early diagnosis of the disease and prevent the manifestation of more severe forms of lesions, helping in the screening process and definition of the best form of treatment. However, the detection of microlesions using computational systems is a challenge due to several factors, such as the size and shape of these lesions, the presence of noise and poor contrast in the images, the small number of labeled examples in public DR datasets, and the difficulty of deep learning algorithms in detecting tiny objects due to gradient dissipation during training. Thus, to overcome these problems, this work proposes two new approaches based on image processing techniques, data augmentation, transfer learning, and deep neural networks to support the medical diagnosis of fundus lesions. We trained, adjusted, and evaluated the proposed approaches using different public Diabetic Retinopathy datasets. We partitioned the datasets into sets of training (50%), validation (20%), and test (30%) to carry out the experiments. We used a validation step to fine-tune the hyperparameters and a test step to assess the generalization capacity of the models. The approach to detecting fundus lesions achieved mAP of 0.2630 for the limit of IoU of 0.5 in the validation step using the DDR dataset and Adam optimizer. The approach for segmenting instances of fundus lesions reached mAP of 0.2903 for the limit of IoU of 0.5 in the validation stage using the DDR dataset and Adam optimizer, thus being 10.38% more accurate than the proposed detection approach. The results obtained in the experiments demonstrate that the new approaches presented promising results in detecting fundus lesions associated with DR.

Keywords: Diabetic Retinopathy. Fundus Images. Deep Learning. Fundus Lesions Detection. Fundus Lesions Instance Segmentation.

LISTA DE FIGURAS

LISTA DE TABELAS

LISTA DE ABREVIATURAS E SIGLAS

2d	2 dimensões
3d	3 dimensões
AG	Algoritmo Genético
AP	*Average Precision*
APTOS	*Asia Pacific Tele-Ophthalmology Society*
AR	*Average Recall*
BF	Filtro Bilateral
AUC	*Area Under The Curve*
BBox	*Bouding Box*
BIBM	*Bioinformatics and Biomedicine*
BN	*Batch Normalization*
CBN	*Cross-Iteration Batch Normalization*
CIE	*International Commission on Illumination*
CLAHE	*Contrast Limited Adaptive Histogram Equalization*
CNN	*Convolutional Neural Network*
COCO	*Common Objects in Context*
CSCI	*Conference on Computational Science and Computational Intelligence*
CSP	*Cross Stage Partial Network*
DDR	*Dataset for Diabetic Retinopathy*
DIARETDB1	*Standard Diabetic Retinopathy Database Calibration level 1*
DL	*Deep Learning*
DLA	*Deep Layer Aggregation*
DM	Diabetes Mellitus
DNNs	*Deep Neural Networks*
DO	Disco Óptico

DPVSA	*Digital Processing of Visual Signals and Applications*
E	Entropia
EMBC	*Engineering in Medicine and Biology Conference*
EMD	Edema Macular Diabético
EME	*Measure of Enhancement*
EX	Exsudato Duro
eyePACS	*Eye Picture Archive Communication System*
Fast R-CNN	*Fast Regions with Convolutional Neural Network features*
Faster R-CNN	*Faster Regions with Convolutional Neural Network features*
FCN	*Fully Convolutional Network*
FC	*Fully Connected*
FN	Falsos Negativos
FP	Falsos Positivos
FPN	*Feature Pyramid Network*
fps	*Frames Per Second*
GFLOPs	*Giga FLoating-point Operations Per Second*
$GIoU$	*Generalized Intersection over Union*
GPU	*Graphics Processing Unit*
HE	Hemorragia
HED	*Holistically-nested Edge Detection*
HOG	*Histogram of Oriented Gradients*
HT	Transformada de Hough
HSV	*Hue, Saturation, Value*
ICO	Internacional Council of Ophthalmology
IDRiD	*Indian Diabetic Retinopathy Image Dataset*
IEEE	Instituto de Engenheiros Eletricistas e Eletrônicos
IJCNN	*International Joint Conference on Neural Networks*
IoU	*Intersection Over Union*
IRMA	*Intraretinal Microvascular Abnormalities*
JPG	*Joint Photographic Experts Group*
JSON	*JavaScript Object Notation*
LAB	Luminosidade, A (coordenada vermelho/verde) e B (coordenada amarelo/azul)
LSTM	*Long Short-Term Memory*

MA	Microaneurisma
mAP	*mean Average Precision*
Mask R-CNN	*Mask Regions with Convolutional Neural Network features*
$mIoU$	*mean Intersection over Union*
MS COCO	*Microsoft Common Objects in Context*
MSE	*Mean Squared Error*
NMS	*Non-max Suppression*
NV	Neovascularização
NVD	Neovasos no disco
NVOL	Neovasos em outros lugares
PAN	*Path Aggregation Network*
PPI	*Pixels Per Inch*
PR	*Precision* $\times$ *Recall*
PSNR	*Peak Signal-To-Noise Ratio*
R-CNN	*Region Based Convolutional Neural Networks*
RD	Retinopatia Diabética
RDNP	Retinopatia Diabética Não-Proliferativa
RDP	Retinopatia Diabética Proliferativa
ReLU	*Rectified Linear*
ResNet	*Residual Neural Network*
RSL	*Revisão Sistemática da Literatura*
RGB	*Red, Green, Blue*
RSL	*Revisão Sistemática da Literatura*
ROC	*Receiver Operating Characteristic*
ROI	*Region of Interest*
ROIAlign	*Region of Interest Align*
RPN	*Region Proposal Networks*
RTX	*Ray Tracing Texel eXtreme*
SAM	*Spatial Attention Module*
SE	Exsudato Algodonoso
SGD	*Stochastic Gradient Descent*
SiLU	*Sigmoid Linear Unit*
SPP	*Spatial Pyramid Pooling*
SSD	*Detector MultiBox Single Shot*

SVM	*Support Vector Machines*
V2	Versão 2
VB	*Venous Beading*
VC	Visão Computacional
VN	Verdadeiros Negativos
VOC	*Visual Objects Classes*
VP	Verdadeiros Positivos
VRAM	*Video Random Access Memory*
XML	*eXtensible Markup Language*
YOLO	*You Only Look Once*

SUMÁRIO

1 INTRODUÇÃO

De acordo com o *ICO Guidelines for Diabetic Eye Care* (2017), a Retinopatia Diabética (RD), uma doença diretamente relacionada ao Diabetes e que afeta os olhos, é uma das principais causas de perda de visão em adultos em idade produtiva. Aproximadamente 1/3 (34,6%) das pessoas com Diabetes nos Estados Unidos da América, Europa e Ásia têm RD. A prevalência do Diabetes está aumentando mundialmente e nas últimas duas décadas a perda parcial ou total da visão cresceu devido às complicações causadas pelo aumento do número de pessoas com Diabetes (VOCATURO; ZUMPANO, 2020).

O Diabetes é uma doença crônica que ocorre quando o pâncreas não produz insulina suficiente (tipo 1), ou quando o corpo não consegue usar com eficácia a insulina que produz (tipo 2). A insulina é um hormônio que regula o açúcar no sangue. A Federação Internacional de Diabetes (em inglês, *International Diabetes Federation* – IDF)[1] relata que, em 2000, o número global estimado de adultos com Diabetes era de 151 milhões. Em 2009 havia crescido 88%, indo para 285 milhões. A IDF estima que haverá 600 milhões de pessoas com Diabetes em 2035 e 700 milhões em 2045 (VOCATURO; ZUMPANO, 2020), conforme apresentado na Figura 1.

De acordo com Vocaturo; Zumpano (2020), a RD tem origem na lesão dos vasos sanguíneos do tecido sensível à luz da retina e trata-se da principal causa de perda de visão em pessoas com idade entre os 20 e 74 anos. Quanto mais tempo uma pessoa tem Diabetes maior será a probabilidade de desenvolver complicações oculares. Para entender a extensão da RD, até 21% dos pacientes com Diabetes tipo 2 têm Retinopatia no momento do primeiro diagnóstico de Diabetes e a maioria desenvolve algum grau de Retinopatia com o tempo. É comum que uma pessoa com Retinopatia Diabética perceba os sintomas somente depois que já foram causados danos à sua visão.

Considerando o número de pacientes afetados por Diabetes no mundo, é importante destacar que para a realização de uma triagem efetiva a fim de verificar as pessoas afetadas pela RD envolve a mobilização de grande quantidade de recursos. A

[1] `https://idf.org/`

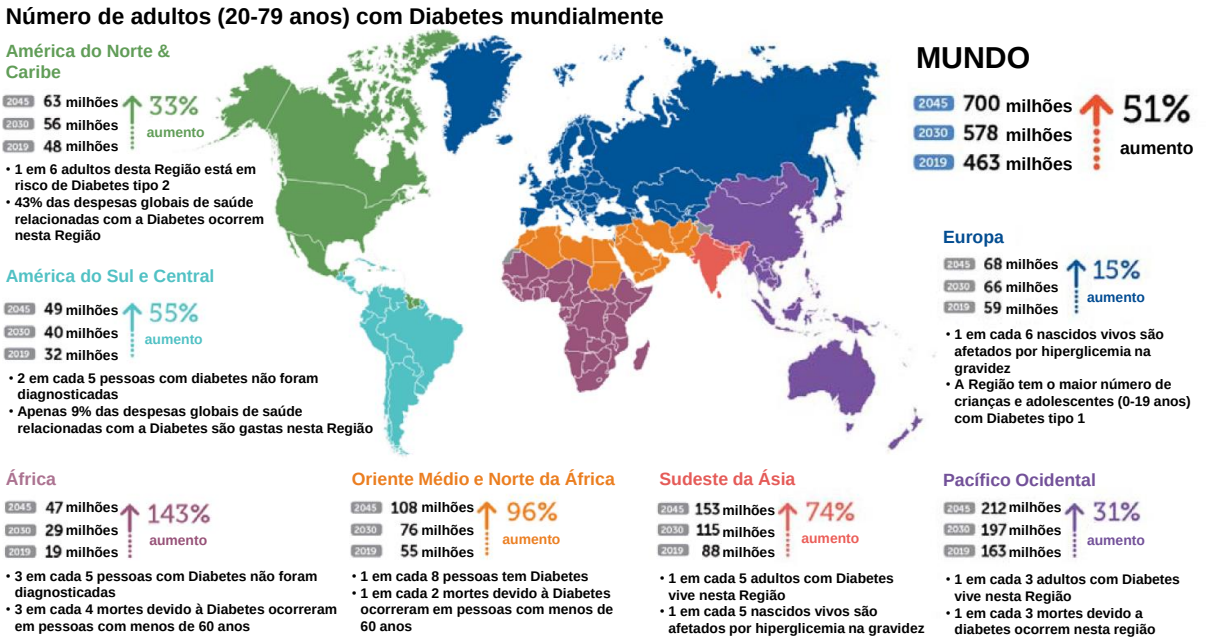


Figura 1 – Relatório com dados sobre a evolução do Diabetes no mundo conforme estudo da Federação Internacional de Diabetes. Fonte: Adaptado de Vocaturo; Zumpano (2020).

maior parte dos problemas na visão causados pela RD são evitáveis com a detecção e tratamento precoce. Embora o método principal para avaliar a RD envolva a oftalmoscopia[2] (*ICO GUIDELINES FOR DIABETIC EYE CARE*, 2017), os grandes volumes de dados e o Aprendizado Profundo (em inglês, *Deep Learning* – DL) podem proporcionar soluções de baixo custo e eficazes para apoiar a realização de diagnósticos precoces, auxiliando o trabalho de profissionais de saúde e permitindo amplo acesso à população a exames para identificação desta doença.

A RD geralmente é identificada por meio de exames oftalmológicos que visam identificar lesões na retina (lesões de fundo) incluindo Exsudatos Duros (EX), Exsudatos Algodonosos (SE), Microaneurismas (MA) e Hemorragias (HE) e pode ser classificada em Retinopatia Diabética Não-Proliferativa (RDNP) ou em Retinopatia Diabética proliferativa (RDP), de acordo com o surgimento de algumas características na retina (NAYAK et al., 2008).

RDNP possui três estágios: inicial, moderado e grave. No estágio inicial, ocorre o aparecimento de MA; no estágio moderado, a retina apresenta MA, HE e SE; e, no estágio severo, há a presença de inúmeros MA, HE e EX. O paciente que está no estágio severo da doença tem grandes chances de em pouco tempo evoluir para RDP, que se caracteriza principalmente pelo aparecimento de neovasos. Neste estágio, os vasos sanguíneos ficam frágeis, correm risco de se romperem e liberarem sangue para o humor vítreo, podendo causar a perda de visão do paciente (FAUST et al., 2012).

[2]Oftalmoscopia é o exame que visa observar a região posterior do globo ocular, que compreende a retina, o disco óptico e os vasos sanguíneos.

A análise de imagens médicas tem grande aplicação no domínio das ciências da saúde, principalmente no auxílio do diagnóstico médico, tanto para prevenção de doenças como para auxílio na escolha do melhor tratamento de doenças. No trabalho apresentado por Ting; Cheung; Wong (2016) foi demonstrado que a perda de visão resultante de RD pode ser evitada quando tratada precocemente. No entanto, a triagem realizada para identificação precoce da RD permanece um desafio, pois diabéticos são geralmente tratados em departamentos de endocrinologia dos hospitais, nos quais costumam-se realizar exames de fundo do olho e este processo geralmente é demorado, pois apenas um número limitado de exames pode ser processado a cada dia. Além disso, o número de oftalmologistas não consegue atender às crescentes demandas ao redor do mundo, particularmente em regiões em desenvolvimento (CHAKRABARTI; HARPER; KEEFFE, 2012).

Os médicos podem identificar a RD por meio da verificação da presença de lesões de fundo associadas às anormalidades vasculares causadas pelo Diabetes. Embora as abordagens convencionais sejam eficazes, suas demandas por recursos são elevadas. O nível de experiência do profissional de saúde que irá atender os pacientes e os equipamentos necessários para a realização dos exames muitas vezes são insuficientes em regiões onde a taxa de Diabetes na população local é alta e a detecção da RD é mais necessária. Além disso, com o número de pessoas com Diabetes crescendo mundialmente, a demanda por infraestrutura para prevenir a perda de visão causada por RD se tornará ainda maior.

Dessa forma, o desenvolvimento de ferramentas computacionais capazes de analisar e processar imagens de fundo podem auxiliar nesse processo de triagem e identificação precoce da doença. Devido ao significado da detecção de lesões em imagem de fundo e à complexidade associada à realização desta tarefa sem o auxílio de um sistema informatizado, muitos métodos automatizados para classificar, segmentar e detectar estas lesões foram desenvolvidos.

O DL tem por propósito permitir que um modelo faça inferências a partir da análise de dados, sendo, portanto, um ramo da inteligência artificial que baseia-se na premissa de que sistemas podem aprender com dados e identificar padrões ocultos. A aprendizagem profunda usa redes neurais artificiais como base. Embora vagamente baseada em redes neurais biológicas, as redes neurais artificiais são uma maneira de especificar um conjunto flexível de funções, construída a partir de muitos blocos computacionais chamados neurônios. Os modelos de aprendizado profundo são treinados com dados do mundo real e aprendem como resolver problemas (ROBERTS; YAIDA; HANIN, 2021).

Soluções apresentadas na literatura auxiliam no diagnóstico de RD por meio de redes neurais profundas, tais como para classificação (LI et al., 2019; LONG et al., 2019; LI et al., 2019; ULLAH et al., 2019; SON et al., 2020; PORWAL et al., 2020;

SHARIF et al., 2020; XU; FENG; MI, 2017; KARKUZHALI; MANIMEGALAI, 2019; THEERA-UMPON et al., 2020; POONKASEM et al., 2019), segmentação (LI et al., 2019; PORWAL et al., 2020; SHARIF et al., 2020; RONNEBERGER; FISCHER; BROX, 2015; CHUDZIK et al., 2018; GUO et al., 2020; YE et al., 2020; IBTEHAZ; RAHMAN, 2020), ou detecção de objetos em imagens (LI et al., 2019; PORWAL et al., 2020; AKYOL; SEN; BAYIR, 2016; PRAKASH; SELVATHI, 2016; PERDOMO; AREVALO; GONZÁLEZ, 2017; AL-MASNI et al., 2018; BENZAMIN; CHAKRABORTY, 2018; QUMMAR et al., 2019; KHOJASTEH et al., 2019; MATEEN et al., 2020; WANG et al., 2020).

Mesmo que redes neurais profundas sejam utilizadas para a detecção de lesões em imagens de fundo, ainda assim há limitações nos resultados obtidos, principalmente provenientes da baixa representatividade dos atributos extraídos das imagens utilizadas para o treinamento dos modelos e da complexidade associada às características como formato, tamanho e incidência destas lesões.

Em função destas limitações, da gravidade da Retinopatia Diabética e, também, do impacto desta doença na saúde e qualidade de vida das pessoas, é oportuno propor a criação e/ou otimização de procedimentos, abordagens e instrumentos computacionais capazes de prover um suporte rápido e preciso ao diagnóstico médico, com o mínimo de intervenção humana. Todos estes aspectos motivam o problema de pesquisa considerado neste trabalho.

1.1 Problema de Pesquisa

O problema de pesquisa central desta Tese consiste em responder a seguinte indagação: *"Como detectar as lesões de fundo associadas à Retinopatia Diabética por meio de redes neurais profundas para o auxílio de diagnóstico médico?"*

Este problema de pesquisa se desdobra nos seguintes aspectos:

(1) como obter um conjunto de dados público que contenha imagens e anotações de lesões de fundo para treinamento de modelos baseados em aprendizado profundo;

(2) como realizar o tratamento das imagens destes conjuntos de dados, a fim de eliminar possíveis ruídos;

(3) como melhorar a qualidade das imagens a fim de permitir uma extração de características mais eficaz das lesões presentes nestas imagens;

(4) como implementar arquiteturas de redes neurais profundas capazes de detectar com precisão as lesões de fundo, levando em conta problemas como o desbalanceamento da quantidade de exemplos das diferentes classes de lesões;

(5) como propor uma abordagem generalizável, que seja capaz de realizar a detecção de lesões de fundo em imagens não conhecidas *a priori*;

(6) como melhorar a precisão da detecção de microlesões de fundo, como no caso dos microaneurismas;

(7) como realizar a detecção e a segmentação de instância de lesões de fundo; e,

(8) como avaliar as abordagens propostas e comparar com soluções análogas encontradas na literatura.

Por fim, é importante considerar que os esforços da pesquisa desenvolvida até o momento tiveram como propósito contribuir com o suporte ao diagnóstico médico no que diz respeito à detecção de lesões associadas à Retinopatia Diabética. Do problema de pesquisa e seus desdobramentos foram consolidados os objetivos, hipóteses e contribuições deste Trabalho, discutidos a seguir.

1.2 Hipóteses

Nesta Tese foram estabelecidas as seguintes hipóteses:

(1) Arquiteturas de aprendizado profundo de estágio único e dois estágios são capazes de detectar de forma eficaz lesões de fundo associadas à Retinopatia Diabética e auxiliar no diagnóstico médico.

(2) Um *pipeline* que inclua etapas para pré-processamento das imagens, aumento de dados e transferência de aprendizado pode aprimorar a detecção das lesões de fundo.

1.3 Objetivos

O objetivo geral desta Tese é propor novas abordagens baseadas em técnicas de processamento digital de imagens e aprendizado profundo para detectar e segmentar instâncias das lesões de fundo associadas à Retinopatia Diabética.

Para alcançar o objetivo geral foram definidos os seguintes objetivos específicos:

(1) proposta de abordagens para a detecção e segmentação de instância de lesões de fundo baseadas em redes neurais profundas;

(2) definição de uma metodologia para preparação, pré-processamento e aumento de dados de imagens de fundo;

(3) definição de abordagens que combinam diferentes técnicas de processamento digital de imagens para melhorar a precisão da detecção e segmentação de instância das lesões de fundo;

(4) estabelecimento de uma arquitetura de rede neural profunda para detecção de objetos ajustada e testada em diferentes conjuntos de dados públicos de RD;

(5) estabelecimento de uma arquitetura de rede neural profunda para segmentação de instância ajustada e testada em diferentes conjuntos de dados públicos de RD; e,

(6) avaliação comparativa das novas abordagens propostas com trabalhos similares encontrados na literatura.

1.4 Organização do Documento

Este documento está organizado como segue. O Capítulo 2 apresenta os Conceitos Básicos sobre Retinopatia Diabética, introduzindo conceitos sobre o Olho Humano, as Lesões de Fundo e a Retinopatia Diabética. No Capítulo 3, são apresentados diferentes Conjuntos de Dados Públicos de Retinopatia Diabética com imagens de fundo para classificação, segmentação e detecção de lesões de fundo.

O Capítulo 4 apresenta modelos de Redes Neurais Profundas para Detecção e Segmentação de objetos em imagens, assim como os conceitos básicos sobre aumento de dados, transferência de aprendizado e métricas para avaliação destes modelos. No Capítulo 5, é apresentada a Revisão da literatura na qual são descritos trabalhos no estado da arte concernentes à identificação de lesões de fundo. Além disso, uma análise comparativa é realizada, em que são elencados os métodos utilizados, as principais contribuições e limitações dos trabalhos selecionados.

O Capítulo 6 apresenta a abordagem proposta para a detecção das lesões de fundo. Neste Capítulo ainda são mostradas as etapas de pré-processamento e aumento de dados, assim como a constituição da arquitetura de rede neural profunda utilizada. Também são apresentados detalhes sobre os materiais e métodos utilizados, os experimentos realizados, e os resultados obtidos pela abordagem proposta.

No Capítulo 7 é apresentada a abordagem proposta para segmentação de instância das lesões de fundo. Neste Capítulo ainda são mostradas as etapas de pré-processamento e aumento de dados, assim como a constituição da arquitetura de rede neural profunda utilizada. Também são apresentados detalhes sobre os materiais e métodos utilizados, os experimentos realizados, e os resultados obtidos pela abordagem proposta.

No Capítulo 8 encontram-se as considerações finais desta Tese, as publicações realizadas, e a sugestão de trabalhos futuros.

2 RETINOPATIA DIABÉTICA

Este Capítulo tem por objetivo apresentar conceitos fundamentais sobre o olho humano, lesões de fundo e Retinopatia Diabética. É importante destacar que o nível de informações oftalmológicas apresentadas está condicionado ao problema de pesquisa abordado e que, portanto, não tem a pretensão de esgotar o assunto.

2.1 Olho Humano

A visão é um dos sentidos mais importantes e complexos, e por meio desta é possível vermos e interagirmos com o mundo ao nosso redor. A visão é baseada na absorção da luz pelas células fotorreceptoras do olho (DELGADO-BONAL; MARTÍN-TORRES, 2016).

Os olhos podem ser prejudicados por diversas doenças. Assim, cuidar deste órgão é fundamental para prevenir ou até mesmo reduzir a severidade destas doenças. A saúde dos olhos está associada a uma melhor qualidade e vida, portanto é necessário que sejam tomados os cuidados necessários para que se mantenha uma visão saudável.

O olho humano é relativamente pequeno, porém extremamente complexo, possuindo um sistema que permite adaptação rápida e sem esforço a ambientes escuros e claros. Dentre as principais estruturas do olho, destacam-se a córnea, o cristalino, a íris, a pupila, a retina, a fóvea, o disco óptico, os vasos sanguíneos e o nervo óptico. A Figura 2 apresenta uma ilustração sumarizada das principais estruturas do olho humano (RICHARDS; MUNAKOMI; MATHEW, 2023).

O globo ocular tem um formato esférico, com raio da ordem de 1,2 cm, sendo em grande parte opaco, com exceção de uma região frontal, onde está a córnea, que é transparente. Após a córnea, há uma lente interna, também chamada de cristalino. A região interna do globo ocular é preenchida por materiais transparentes: entre a córnea e a lente há um líquido chamado de humor aquoso; e, depois da lente, o globo ocular é preenchido pelo humor vítreo (HELENE; HELENE, 2011).

Internamente, logo após a córnea, há uma pequena abertura designada por pupila,

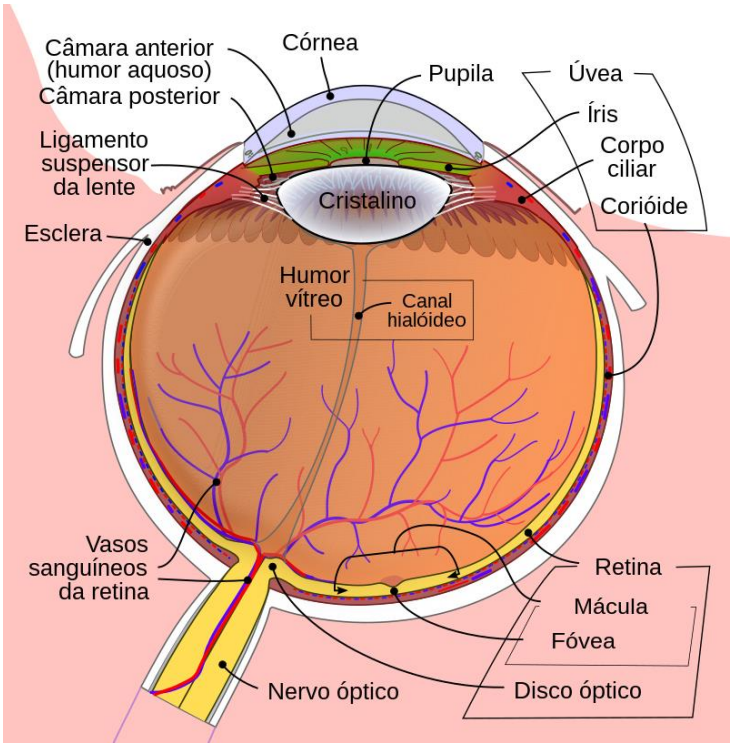


Figura 2 – Ilustração das principais estruturas do Olho Humano. Fonte: Adaptado de Richards; Munakomi; Mathew (2023).

por onde a luz penetra. A pupila tem diâmetro variável, entre 2 mm e 6 mm, dependendo de fatores como iluminação, por exemplo. A abertura da pupila é controlada pelos músculos da íris, que é uma estrutura circular cuja cor da parte externa pode variar, sendo as mais comuns o marrom, o azul e o verde.

A retina possui uma região muito densa de células sensíveis à luz designada por fóvea, que fica na direção frontal do olho, ao longo de seu eixo principal. Do ponto de vista do diagnóstico médico, a análise da imagem da retina (imagem de fundo) é uma abordagem fundamental para identificar doenças oculares.

A córnea, tem um índice de refração na ordem de 1,38, e um raio de curvatura de aproximadamente 0,80 cm na sua parte anterior e de cerca de 0,65 cm na parte posterior. Sua espessura é de cerca de 0,06 cm na parte central (o polo, sobre o eixo principal) e um pouco maior na parte lateral (HELENE; HELENE, 2011).

Segundo Helene; Helene (2011), a lente interna ou cristalino tem raios de curvatura que podem variar, permitindo focar imagens mais próximas ou mais distantes. A espessura da lente é de aproximadamente 0,4 cm, sendo que a distância entre a superfície anterior da lente e a córnea é cerca de 0,35 cm. O índice de refração da lente não é uniforme, podendo variar do centro para a borda, porém, com valor aproximado de 1,42. No entanto, convém destacar que todas estas dimensões geométricas podem variar entre as pessoas, sendo estes valores aproximativos.

A seguir serão apresentadas com maior nível de detalhe as estruturas do olho que são importantes para compreensão deste trabalho.

2.1.1 Disco Óptico

O Disco Óptico (DO) aparece como uma região oval amarelada brilhante nas imagens coloridas de fundo, através do qual os vasos sanguíneos entram no olho (AL-BANDER et al., 2018), considerado o local para onde as fibras nervosas do olho convergem, de onde sai o nervo óptico e entram os vasos sanguíneos (BESENCZI; TÓTH; HAJDU, 2016). É uma estrutura arredondada ou elipsoide, com cerca de 1,5 mm que corresponde a cabeça do nervo óptico, possuindo cor rosa a laranja com uma depressão amarelo-pálida no centro (GAGNON et al., 2001).

O DO é frequentemente considerado como um ponto de referência para localizar outras estruturas retinianas. Por exemplo, pode ser usado como ponto de partida para rastreamento de vasos retinianos em algoritmos de rastreamento de vasos sanguíneos (AL-BANDER et al., 2018).

2.1.2 Vasos Sanguíneos

O olho é suprido pela artéria oftálmica, que é o primeiro ramo da artéria carótida interna, ao passar pelo seio cavernoso. A artéria oftálmica possui numerosas ramificações que suprem os músculos que movem o olho e circundam o olho, a pálpebra e o próprio globo ocular. Os ramos da artéria oftálmica são divididos em orbitais (suprem a órbita e estruturas relacionadas) e o grupo óptico (suprem o olho e seus músculos). A retina é irrigada pelos seguintes sistemas vasculares (MAIA, 2018):

Artéria central da retina: a artéria central da retina se apresenta ao interior do globo ocular pela papila óptica. A partir de sua primeira bifurcação os vasos já são arteríolas, perfeitamente visíveis à fundoscopia. Os ramos da artéria central da retina percorrem a superfície retiniana e são responsáveis pelo suprimento sanguíneo para os 2/3 internos da retina.

Artéria retiniana central: passa por baixo do nervo óptico e envia vários ramos sobre o aspecto interno da retina. Na Retinopatia Diabética pode haver a formação de hemorragias e aneurismas que podem se formar nesta artéria e nos seus ramos (LANDAU; KURZ-LEVIN, 2011).

Coriocapilares: representados pelas artérias ciliares posteriores (medial e lateral), constituem-se um ramo da artéria oftálmica. As camadas retinianas externas são avasculares e esses vasos são responsáveis por suprir 1/3 externo da espessura da retina pela difusão coriocapilar (epitélio pigmentar da retina, camada dos fotorreceptores e camada nuclear interna) (MAIA, 2018).

Veia central da retina: a drenagem venosa retiniana geralmente acompanha o suprimento arterial visto na fundoscopia. Existem pontos de cruzamento, principalmente as vênulas presentes na retina interna, que entrelaçam com as arteríolas

associadas. Esses pontos são locais mais comuns de obstrução do ramo da veia retiniana (MAIA, 2018).

2.1.3 Mácula

A região mais posterior e central da retina é denominada mácula, conforme ilustrado na Figura 3, apresentando aspecto ovalado. A margem macular tem diâmetro aproximado de 5,5 mm. A mácula pode ser diferenciada da retina periférica pela grande camada de células ganglionares, que apresentam maior espessura, enquanto a periferia retiniana tem a espessura de uma só célula (MAIA, 2018).

A mácula pode ser observada no centro da retina como uma mancha redonda mais escura, cujo centro é conhecido como fóvea, que é responsável pela nitidez da visão (BESENCZI; TÓTH; HAJDU, 2016).

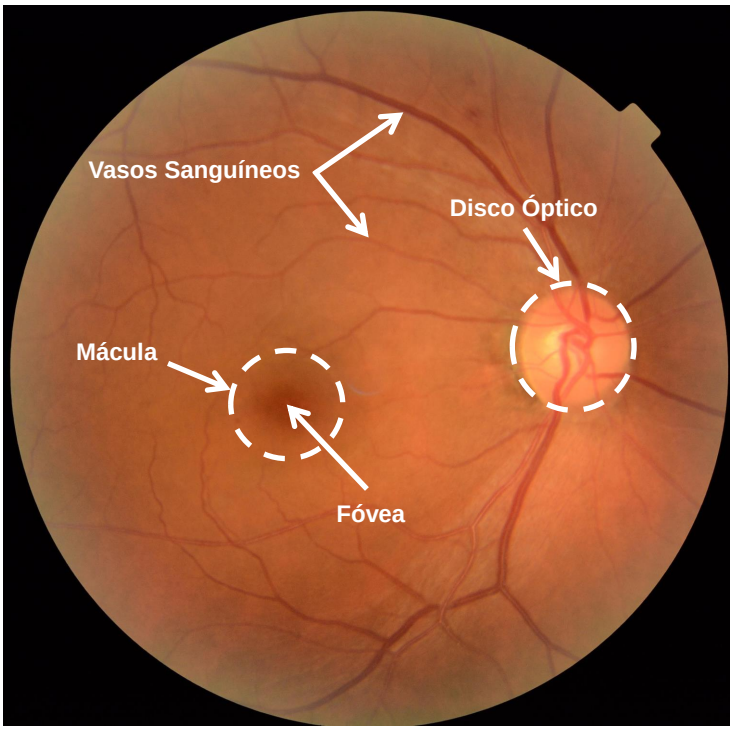


Figura 3 – Imagem de fundo com destaque para a mácula e a fóvea (parte central), disco óptico (à direita), e os vasos sanguíneos (parte superior).

2.1.4 Fóvea

A fóvea está localizada na mácula, sendo uma área importante da retina por concentrar grande parte das células fotorreceptoras. Pode ser observada no lado temporal do disco óptico, a uma distância de aproximadamente 2,5 vezes o diâmetro do DO (CARDOSO, 2019). A Figura 3 apresenta um exemplo de imagem de fundo, na qual a mácula e a fóvea são mostradas na parte mais central à esquerda, o disco óptico à direita e os vasos sanguíneos, na parte superior, entrando e saindo do olho através do disco óptico.

A fóvea possui três características discriminantes: 1) É populada apenas por cones[1], não possuindo bastonetes[2], o que lhe confere alta capacidade para identificação de cores; 2) Nesta região, cada célula ganglionar formadora do nervo óptico se liga a uma única célula fotossensível, o que lhe confere alta resolução para a imagem que será enviada ao córtex visual; e, 3) Há um deslocamento lateral das células não fotossensíveis da retina, não havendo formação de sombras sobre os cones, aumentando sua sensibilidade e lhe conferindo o formato abaulado da retina nessa região. O diâmetro da fóvea é da ordem de 1 mm e nela são projetadas as imagens que somos capazes de distinguir com precisão (HELENE; HELENE, 2011).

2.1.5 Retina

A retina humana é o mais complexo dos tecidos oculares, possuindo uma estrutura altamente organizada. A retina recebe a imagem visual, produzida pelo sistema óptico do olho, e converte a energia da luz em um sinal elétrico, que após um processamento inicial é transmitido através do nervo óptico para o córtex visual (RIORDAN-EVA; AUGSBURGER, 2018). É uma camada fina de tecido nervoso, semitransparente e com multicamadas que reveste a parte interna dos dois terços posteriores da parede do globo ocular.

É sensível à luz e pode ser comparada a um filme numa câmera fotográfica, sendo como uma tela para projetar as imagens enxergadas, que retém as imagens, traduzindo para o cérebro através dos impulsos elétricos enviados pelo nervo óptico ao cérebro (IORJ, 2021).

Em pacientes com Diabetes a retina pode ser afetada pela patologia conhecida como Retinopatia Diabética (MOOKIAH et al., 2013; RIORDAN-EVA; AUGSBURGER, 2018), que ocorre quando um material anormal é depositado nas paredes dos vasos sanguíneos da retina, que é a região conhecida como fundo do olho, causando o estreitamento e por vezes o bloqueio do vaso sanguíneo, além de enfraquecimento da parede do vaso e o surgimento de lesões de fundo (IORJ, 2021), que serão abordadas na próxima Seção.

2.1.6 Exame de Fundo do Olho

O exame de fundo do olho proporciona a análise da retina e de seus elementos, sendo um exame de alta importância para a identificação de diferentes patologias oftalmológicas que afetam o fundo do olho (DORION, 2002). Com o exame de fundo

[1] Células fotossensíveis menos sensíveis à luz. Estas células conseguem captar diferentes comprimentos de onda, permitindo a visão em cores. Em cada olho humano é possível encontrar cerca de 6 milhões de cones, sendo a maioria concentrada na fóvea, região onde a imagem se forma com maior nitidez.

[2] Células fotossensíveis capazes de captar imagens mesmo com pouca luminosidade, sendo extremamente sensíveis à luz. Essas células são incapazes de distinguir cores. Em cada olho humano é possível encontrar cerca de 120 milhões de bastonetes.

do olho é possível identificar diversas doenças degenerativas, metabólicas, genéticas, inflamatórias e infecciosas sistêmicas.

Na retina doente podem existir alterações oculares (lesões de fundo), tais como:

Exsudatos duros: que são lesões amarelo-esbranquiçadas, intra-retinianas, resultado do extravasamento de lipídios, que se acumulam nas camadas mais profundas da retina;

Exsudatos algodonosos: que são a consequência de isquemias, geralmente localizados nas camadas mais superficiais; e,

Hemorragias: que são pontos arredondados, situados nas camadas mais profundas da retina (CARDOSO, 2019).

Estas alterações oculares serão abordadas com maior detalhamento na Seção seguinte.

2.2 Lesões de Fundo

Além dos elementos presentes na retina, diversas lesões podem ser observadas durante um exame de fundo do olho. A seguir serão apresentadas as principais lesões de fundo, conforme Dorion (2002); Mookiah et al. (2013); Hendrick; Gibson; Kulshreshtha (2015); Pires; Rocha; Wainer (2019).

2.2.1 Microaneurismas

São pequenas lesões na forma de pequenos pontos circulares na cor vermelha que aparecem na retina. São lesões em formato sacular causados por um vazamento proveniente de uma fraqueza vascular em um determinado ponto. Representam os primeiros sinais de RD (HENDRICK; GIBSON; KULSHRESHTHA, 2015; WILLIAMS et al., 2004). A permeabilidade anormal e a não perfusão do sangue nos capilares sanguíneos causa a formação de MA (WILLIAMS et al., 2004). É uma mancha vermelha menor que 125 μ de tamanho e tem margens nítidas (ETDRSR, 1991a).

Segundo o *ICO Guidelines for Diabetic Eye Care* (2017), microaneurismas são pontos vermelhos, esféricos e isolados de tamanhos variados. Eles podem refletir uma tentativa abortiva de formar um novo vaso ou podem simplesmente ser uma fraqueza da parede do vaso capilar devido à perda da integridade estrutural normal. A Figura 4 apresenta um exemplo de microaneurisma em uma imagem de fundo.

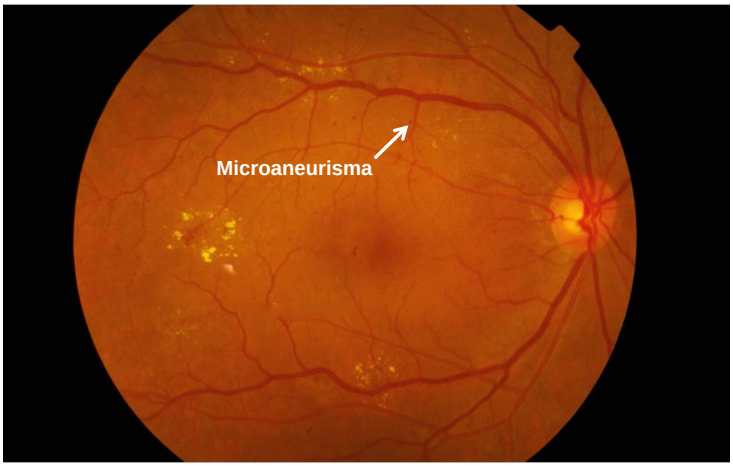


Figura 4 – Retinopatia diabética Não-Proliferativa moderada com hemorragias, exsudatos duros e microaneurismas. Fonte: Adaptado de *ICO Guidelines for Diabetic Eye Care* (2017).

2.2.2 Hemorragias

Estes sinais podem ocorrer na camada pré-retiniana, na retina, ou no vítreo. Seu aspecto pode ser diferente de acordo com o local em que ocorre. Hemorragias pré-retinianas costumam ser volumosas e se acumulam em bolsas de formato arredondado semelhante a uma taça, e não são provenientes de RD (CARDOSO, 2019).

As hemorragias intra-retinianas são lesões vermelhas maiores e mais irregulares na retina. Possuem um aspecto intimidador, porém geralmente desaparecem entre 3 a 4 meses. Ocorrem devido a vazamento em vasos capilares fracos (HENDRICK; GIBSON; KULSHRESHTHA, 2015). É definido como uma mancha vermelha com margem e/ou densidade irregular. Normalmente seu tamanho é maior que 125 μ (ETDRSR, 1991a). A Figura 5 apresenta um exemplo de hemorragia em uma imagem de fundo.

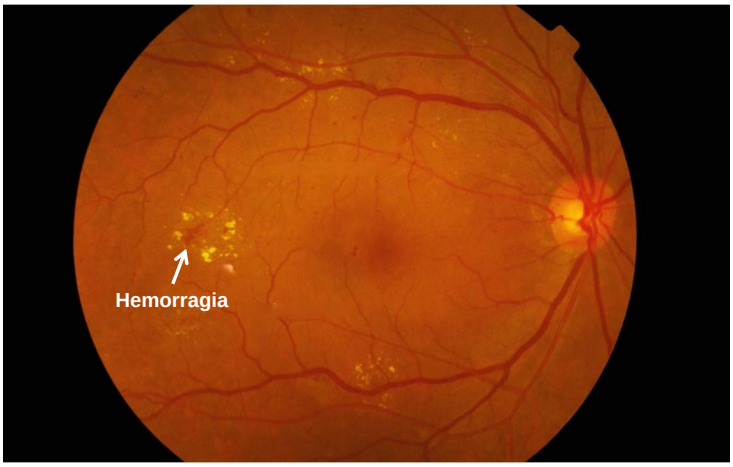


Figura 5 – Retinopatia Diabética Não-Proliferativa moderada com hemorragias, exsudatos duros e microaneurismas. Fonte: Adaptado de *ICO Guidelines for Diabetic Eye Care* (2017).

2.2.3 Exsudatos Duros

São as lipoproteínas e outras proteínas que vazam pelos vasos retinianos anormais (ALGHADYAN, 2011). São lesões amarelas de forma irregular e, quando acompanhado de espessamento da retina, representam uma característica do Edema Macular Diabético (EMD) (HENDRICK; GIBSON; KULSHRESHTHA, 2015).

Possuem também cor amarela-esbranquiçada, com aspecto brilhante e margens bem delimitadas (ETDRSR, 1991a). Em exames de retinografia possuem um aspecto refrativo e podem aparecer em diversas disposições: isolados, em cachos, em trilhas confluentes ou em anel parcial ou completo (CARDOSO, 2019). A Figura 6 apresenta exemplos de exsudatos duros em uma imagem de fundo.

Frequentemente, se localizam no polo posterior, perto ou ao redor da mácula. Podem configurar alguns pontos ou até formarem placas, alguns se dispersam em 4 a 6 meses e outros permanecem por anos (ETDRSR, 1991a).

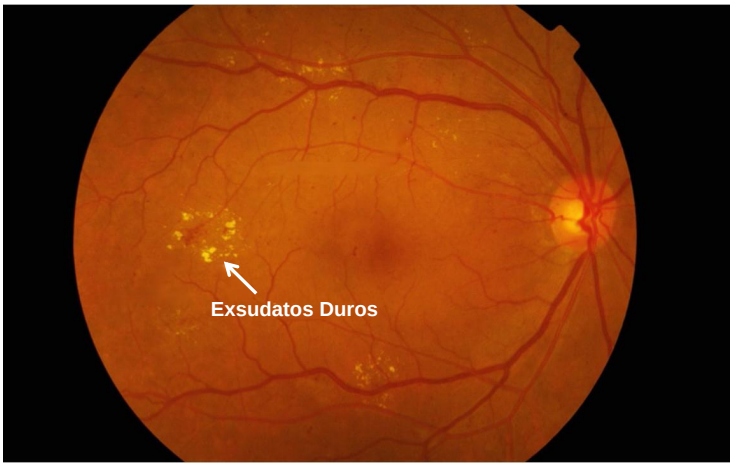


Figura 6 – Retinopatia Diabética Não-Proliferativa moderada com hemorragias, exsudatos duros e microaneurismas. Fonte: Adaptado de *ICO Guidelines for Diabetic Eye Care* (2017).

2.2.4 Exsudatos Algodonosos

Estes sinais conhecidos como exsudatos algodonosos ou moles são microinfartos retinianos resultantes de doenças como Diabetes, hipertensão, oclusão da veia retiniana, papiledema, doenças do colágeno, anemia, leucemia e síndromes de hiperviscosidade (CARDOSO, 2019).

Ocorrem devido à oclusão da arteríola (MCLEOD, 2005) e a redução do fluxo de sangue para a retina causa isquemia da camada de fibra nervosa retinal, afetando o fluxo axoplasmático e acumulando detritos axoplasmáticos nos axônios das células ganglionares da retina (MOOKIAH et al., 2013). Esta lesão aparece em forma de manchas e possuem aspecto pálido, geralmente em tons branco-acinzentados, possuindo margens borradas e irregulares, costumeiramente encontradas em volta do disco óptico, como apresenta a Figura 7.

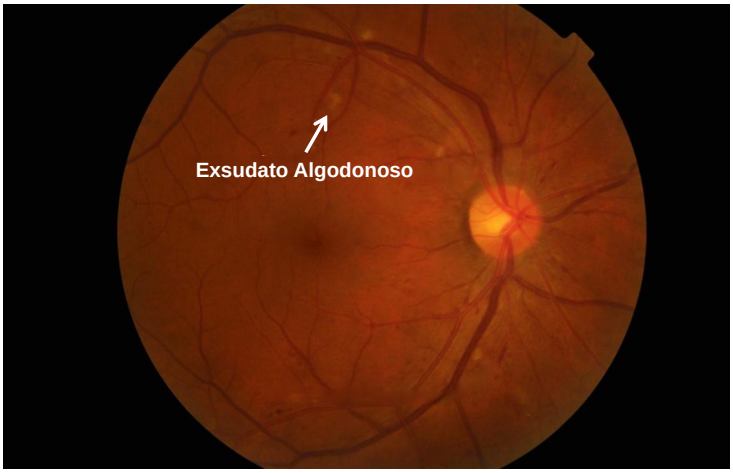


Figura 7 – Retinopatia Diabética Não-Proliferativa moderada com edema macular diabético. Fonte: Adaptado de *ICO Guidelines for Diabetic Eye Care* (2017).

2.3 Retinopatia Diabética

De acordo Riordan-eva; Augsburger (2018), a prevalência global de Diabetes Mellitus (DM) em adultos é de cerca de 8,5%, com um aumento de quatro vezes em quantidade de pessoas com Diabetes entre 1980 e 2014. O DM é um distúrbio metabólico complexo que pode levar à doença vascular generalizada, sendo caracterizado pelo comprometimento do metabolismo da glicose causado pela deficiência de insulina ou sua ausência, levando à hiperglicemia, que pode finalmente resultar em complicações vasculares e neuropáticas (MOOKIAH et al., 2013).

O risco de complicações oculares é aumentado pela falta de controle do Diabetes e da hipertensão sistêmica, porém, mesmo com um monitoramento adequado destas doenças ainda sim podem ocorrer complicações. O aumento da expectativa de vida dos diabéticos resulta em um aumento acentuado na prevalência de complicações oculares, cujo prognóstico é geralmente melhor para diabéticos tipo 2 do que para diabéticos tipo 1 (RIORDAN-EVA; AUGSBURGER, 2018).

Os níveis de insulina precisam ser monitorados constantemente e a falha de um bom controle glicêmico pode levar a danos dos órgãos, incluindo a Retinopatia Diabética, que pode resultar em perda de visão (HARNEY, 2006; RIORDAN-EVA; AUGSBURGER, 2018). Todos os diabéticos podem eventualmente desenvolver RD (ALGHADYAN, 2011). Pacientes com Retinopatia Diabética Proliferativa apresentam risco elevado de insuficiência ou ataque cardíaco, acidente vascular cerebral, nefropatia diabética, amputação e morte (HARNEY, 2006; AHMAD FADZIL et al., 2011).

A RD é um distúrbio vascular da retina que ocorre frequentemente no DM. A maioria das alterações envolve a circulação venosa (as veias são vasos sanguíneos responsáveis pela condução do sangue dos tecidos periféricos para o coração). Como conduzem sangue a baixa pressão não necessitam de paredes tão resistentes, porém precisam de válvulas para evitar o refluxo do sangue.

Este distúrbio pode ser causado pela perda de perícitos[3] intramurais dos capilares, seguido por perfuração da parede capilar e depois por uma permeabilidade anormal dos capilares, os quais podem levar ao edema retiniano. Alternativamente, RD pode ser causado por um fechamento capilar progressivo elaborando um fator vasoproliferativo que pode estimular a neovascularização retiniana. Por último, o tecido fibrovascular pode contrair-se, produzindo hemorragia vítrea e descolamento retiniano tradicional (DORION, 2002).

Nos Estados Unidos, entre 40 e 45% dos diabéticos têm Retinopatia. Os diabéticos tipo 1 desenvolvem uma forma grave de Retinopatia em 20 anos em cerca de 60 a 75% dos casos, mesmo com um bom controle do Diabetes. Nos pacientes com Diabetes tipo 2, geralmente mais velhos, a Retinopatia mais frequente é a não proliferativa (RIORDAN-EVA; AUGSBURGER, 2018). Os estágios iniciais da RD podem ser clinicamente assintomáticos e se a doença for detectada em estágios avançados o tratamento pode se tornar difícil (YEN; LEONG, 2008).

De acordo com Riordan-eva; Augsburger (2018), é importante que o tratamento de RD seja administrado antes que ocorra a perda da visão, e para que isto ocorra os diabéticos devem fazer exame de fundo regularmente, e serem encaminhados sempre que houver indicação de tratamento. Mais importante que a prevenção da Retinopatia Diabética, é o controle do nível de açúcar no sangue, da pressão arterial, dos lipídios séricos e da função renal.

Os programas de triagem geralmente dependem da revisão de pelo menos fotografias anuais de fundo após a dilatação da pupila, com encaminhamento a um oftalmologista quando são detectadas anormalidades que ameacem a visão. É necessária uma triagem mais frequente durante a gravidez. Qualquer Diabetes que desenvolva perda visual deve ser encaminhado para avaliação oftálmica (RIORDAN-EVA; AUGSBURGER, 2018).

A detecção precoce e o tratamento no momento mais adequado da RD é essencial para a prevenção da perda visual permanente. O rastreamento deve ser realizado dentro de 3 anos a partir do diagnóstico de Diabetes tipo 1 e no momento do diagnóstico no caso de Diabetes tipo 2. A partir disso, anualmente, em ambos os tipos. A Retinopatia Diabética pode progredir rapidamente durante a gravidez e o rastreamento deve ser realizado no primeiro trimestre e, na sequência, pelo menos a cada 3 meses até o parto (RIORDAN-EVA; AUGSBURGER, 2018). Avanços recentes em imagens, particularmente quanto à fotografia de fundo, melhoraram a detecção da Retinopatia, sendo portanto um método eficaz de triagem. Dependendo da presença de características clínicas, a RD é classificada como (ALGHADYAN, 2011; ETDRSR, 1991a; PHILIP

[3]Os perícitos ajudam na regeneração dos vasos quando são rompidos e, além disso, fazem uma contração para evitar a perda sanguínea. Perícitos estão presentes nas vênulas, que são veias de pequeno calibre (0,2 a 1 mm), que estabelecem a ligação entre os capilares e as veias de maior calibre.

et al., 2007; ETDRSR, 1991b): Retinopatia Diabética Não-Proliferativa (RDNP) leve; RDNP moderada; RDNP grave; Retinopatia Diabética Proliferativa (RDP); e, Edema Macular Diabético (EMD). Mais detalhes sobre as características da Retinopatia Diabética podem ser consultados nos Anexos A, B e C.

2.4 Considerações sobre o Capítulo

Este Capítulo apresentou conceitos básicos para a compreensão do trabalho. Dentre os principais, destacam-se a descrição das principais estruturas do olho humano, com destaque para o Disco Óptico, Vasos Sanguíneos, Mácula, Fóvea e Retina.

Também foi discutida a importância da realização do exame do fundo de olho para identificação das diferentes patologias oftalmológicas, tais como doenças degenerativas, metabólicas, genéticas, inflamatórias e infecciosas sistêmicas.

Por fim, foram apresentados os conceitos relacionados às lesões microvasculares da retina, tais como Microaneurismas, Hemorragias, Exsudatos Duros, Exsudatos Algodonosos, e as caraterísticas clínicas para identificação destas lesões.

No próximo Capítulo serão apresentados os conjuntos de dados públicos de Retinopatia Diabética e, também, discutidas as características destas bases de imagens de fundo.

3 CONJUNTOS DE DADOS PÚBLICOS DE RETINOPATIA DIABÉTICA

Este Capítulo tem por objetivo apresentar conjuntos de dados públicos de Retinopatia Diabética, destacando suas características e peculiaridades, com o propósito de selecionar os conjuntos de dados mais adequados para a realização dos experimentos propostos para o desenvolvimento da pesquisa.

Há diferentes conjuntos de dados disponibilizados publicamente por pesquisadores e centros de pesquisa em oftalmologia. Estas bases de dados contêm imagens de retinografias digitais que permitem a realização de pesquisas relacionadas à classificação de RD, segmentação e detecção de objetos, incluindo lesões de fundo associadas à Retinopatia Diabética. Citam-se como exemplos de bases públicas: DDR[1], IDRiD[2], Kaggle eyePACS[3], Kaggle APTOS 2019[4], e-Ophtha[5], DIARETDB1[6], Messidor[7], HEI-MED[8], DRIVE[9], STARE[10], HRF[11].

Os conjuntos de dados têm diferentes características entre si, como a quantidade e a qualidade das imagens, o método de anotação e a quantidade de anotações das lesões, a disponibilização de *Ground Truth*, etc. Para realizar a seleção dos conjuntos de dados mais adequados para a realização da pesquisa foram levados em conta critérios como a quantidade e a qualidade das imagens disponíveis nestes conjuntos de dados, assim como a disponibilização de anotações em nível de *pixel* e o fornecimento de *Ground Truth* das lesões anotadas, conforme apresentado na Tabela 1. A seguir, são apresentados alguns dos principais conjuntos de dados públicos de RD.

[1] https://github.com/nkicsl/DDR-dataset
[2] https://ieee-dataport.org/open-access/indian-diabetic-retinopathy-image-dataset-idrid
[3] https://www.kaggle.com/c/diabetic-retinopathy-detection
[4] https://www.kaggle.com/c/aptos2019-blindness-detection
[5] https://www.adcis.net/en/third-party/e-ophtha
[6] https://www.it.lut.fi/project/imageret/diaretdb1
[7] https://www.adcis.net/en/third-party/messidor/
[8] https://github.com/lgiancaUTH/HEI-MED
[9] https://drive.grand-challenge.org/
[10] https://cecas.clemson.edu/~ahoover/stare/
[11] https://www5.cs.fau.de/research/data/fundus-images/

3.1 *Dataset for Diabetic Retinopathy*

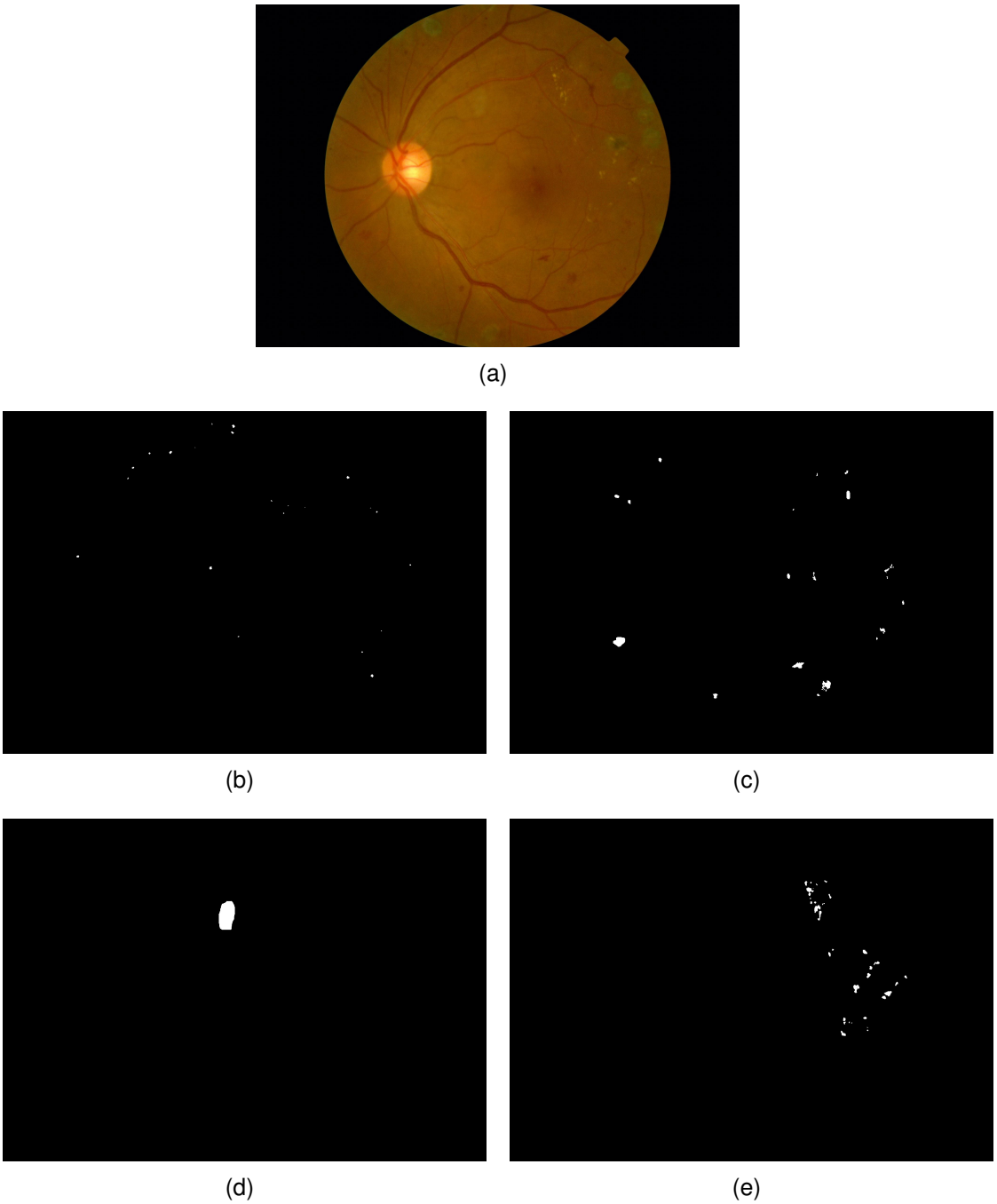
Para a construção do conjunto de dados público de Retinopatia Diabética DDR (em inglês, *Dataset for Diabetic Retinopathy*), foram coletadas 13.673 imagens coloridas do fundo do olho de 147 hospitais entre 2016 a 2018, cobrindo 23 províncias na China. Estas imagens foram fornecidas por 9.598 pacientes com idades entre 1 e 100 anos, sendo a média de idade igual a 54,13 anos (LI et al., 2019).

Dentre essas imagens, 48,23% são do sexo masculino pacientes e 51,77% são de pacientes do sexo feminino. As imagens de fundo foram capturadas usando 42 tipos de câmeras de fundo com um FOV de 45°, principalmente Topcon D7000, Topcon TRC NW48, Nikon D5200 e Canon CR 2 (LI et al., 2019). O conjunto de dados fornece três tipos de anotações: anotações de classificação de RD em nível de imagem, anotações em nível de *pixel* para segmentação de lesões associadas à RD e anotações de caixa delimitadora das lesões associadas à RD.

Segundo Li et al. (2019), todas as imagens foram classificadas para determinar a gravidade da RD por oftalmologistas de acordo com a Classificação Internacional de Retinopatia Diabética (OPHTHALMOLOGY, 2017). As imagens do conjunto de dados DDR são divididas em seis classes: sem RD, RDNP leve, RDNP moderada, RDNP grave, RDP e não classificável. Graduadores profissionais foram treinados por oftalmologistas do Hospital Beijing Tongren e do Instituto de Oftalmologia de Pequim.

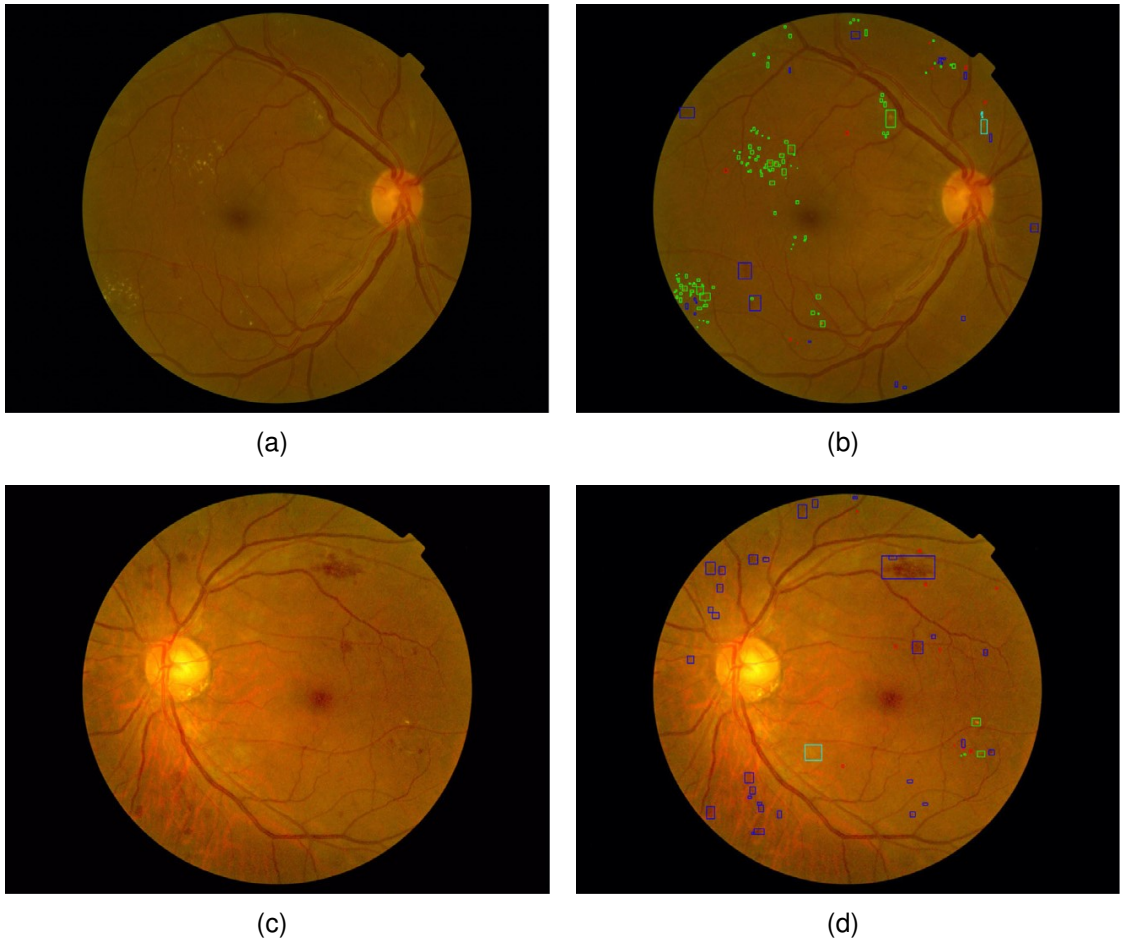
A Figura 8 apresenta um exemplo de imagem original e as anotações em nível de *pixel* das lesões de fundo. As anotações de caixa delimitadora foram geradas automaticamente a partir das anotações em nível de *pixel*. A Figura 9 mostra exemplos de imagens com anotações de caixa delimitadora das lesões de fundo.

Para garantir que os dados de treinamento e as distribuições de dados de teste fossem aproximadamente correspondentes, foram selecionadas aleatoriamente 50% das imagens como um conjunto de treinamento, 20% como um conjunto de validação, e 30% como um conjunto de teste. Além disso, este conjunto de dados fornece as imagens originais com as anotações e *Ground Truth* das lesões de fundo (LI et al., 2019).



(a) (b) (c) (d) (e)

Figura 8 – Exemplo de imagem de fundo do *dataset* DDR com as anotações em nível de *pixel* de Microaneurismas, Hemorragias, Exsudatos Algodonosos e Exsudatos Duros.



(a) (b)

(c) (d)

Figura 9 – Imagens de amostra com anotações de caixa delimitadora das lesões de fundo. As imagens originais estão à esquerda, e as imagens anotadas à direita. As anotações em vermelho representam MA; as anotações em azul representam HE; as anotações em azul claro representam SE; e, as anotações em verde representam EX. Fonte: Adaptado de Li et al. (2019).

3.2 *Indian Diabetic Retinopathy Image Dataset*

O conjunto de dados público IDRiD (em inglês, *Indian Diabetic Retinopathy Image Dataset*), foi criado a partir de exames clínicos adquiridos em uma clínica de olhos localizada em Nanded, Índia. Fotografias da retina de pessoas afetadas por Diabetes foram capturadas com foco na mácula usando a câmera de fundo do olho Kowa VX-10 com um FOV de 50° (PORWAL et al., 2020). O conjunto de dados final é composto por 516 imagens com resolução de 4.288×2.848 *pixels* e são armazenadas no formato de arquivo JPG (em inglês, *Joint Photographic Experts Group*).

Para a construção do conjunto de dados, os especialistas médicos avaliaram o nível de gravidade de RD em uma escala de 0 (sem RD) a 4 (RD grave) e o risco de edema macular diabético em uma escala de 0 (sem EMD) a 2 (EMD grave), para o conjunto completo de 516 imagens. O conjunto de dados de classificação de doenças foi dividido em um conjunto de treinamento, com 413 imagens (80%), e um conjunto de teste, com 103 imagens (20%).

Em seguida, foram realizadas anotações em nível de *pixel* das lesões típicas associadas à RD, e também do disco óptico. Foram disponibilizadas 81 imagens com anotações de lesões de fundo. Junto com estas anotações, o conjunto de dados também fornece as localizações em nível de *pixels* do disco óptico e da fóvea para todas as imagens. A Figura 10 apresenta um exemplo de imagem de fundo (a) original e os *Ground Truth* para (b) MA, (c) HE, (d) SE, (e) EX e (f) DO, respectivamente.

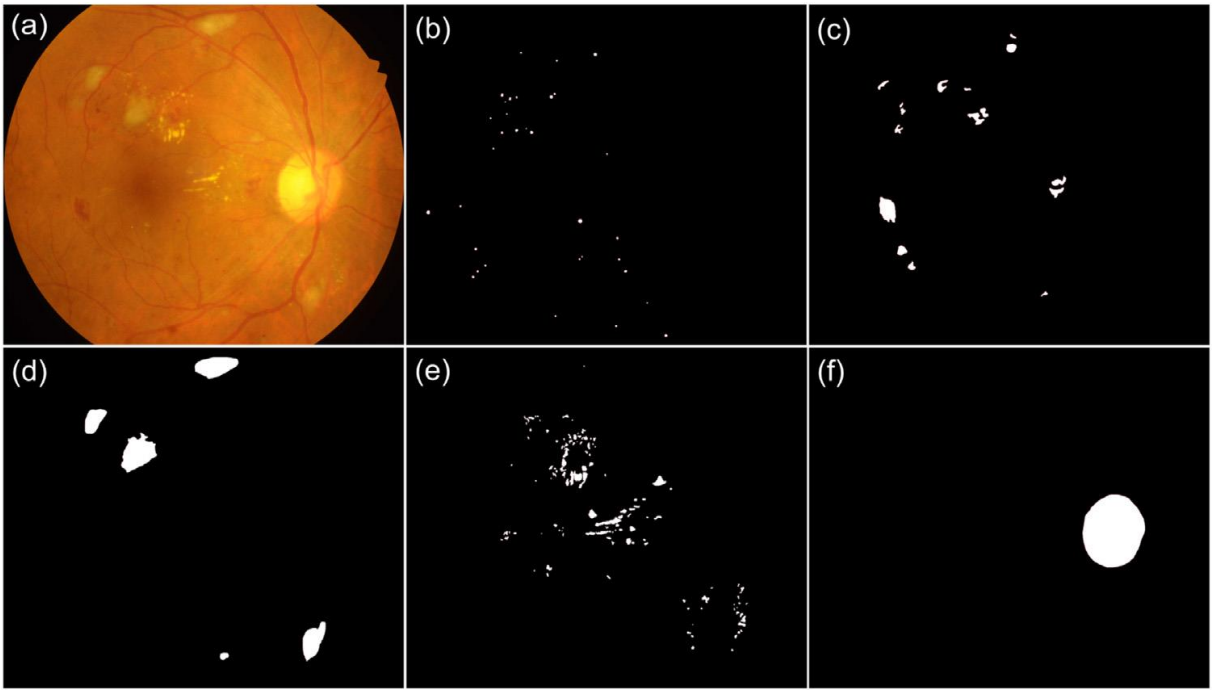


Figura 10 – Exemplo de imagem de fundo do *dataset* IDRiD com as anotações de Microaneurismas, Hemorragias, Exsudatos Algodonosos, Exsudatos Duros e Disco Óptico. Fonte: (PORWAL et al., 2020).

IDRiD é um conjunto de dados de alta qualidade, que além de fornecer patologias associadas à RD e EMD disponibiliza as anotações das principais estruturas de fundo. Uma desvantagem desse conjunto de dados é a pequena quantidade de imagens com anotações das lesões. O conjunto de dados IDRiD está disponível no IEEE Dataport Repository[12].

Tabela 1 – Comparação entre os conjuntos de dados públicos de retinopatia diabética utilizados nesta Tese.

Conjunto de Dados	Total de Imagens	Resolução	Quantidade de Imagens com Anotações por tipo de Lesão				Anotações em nível de *pixel*	Múltiplos Especialistas	Tarefa
			MA	HE	EX	SE			
DDR	12.522	Variável	570	601	486	239	Sim	Sim	Classificação de RD Segmentação de Lesões Detecção de Lesões
IDRiD	516	4288×2848	81	80	81	40	Sim	Sim	Classificação de RD Segmentação de Lesões

Em comparação com os demais conjuntos de dados disponíveis atualmente, o *dataset* DDR apresenta a melhor relação entre quantidade e qualidade de imagens disponíveis. No DDR as lesões foram anotadas por múltiplos especialistas, o que proporciona maior confiabilidade ao conjunto de dados.

O conjunto de dados IDRiD tipicamente foi disponibilizado para tarefas de classificação de RD e segmentação de lesões, enquanto que o *dataset* DDR pode ser utilizado para classificação de RD, segmentação e detecção de lesões de fundo. Dentre os conjuntos de dados analisados, somente os *datasets* DDR e IDRiD possuem anotações em nível de *pixel* das lesões associadas à RD.

3.3 Considerações sobre o Capítulo

Este Capítulo apresentou os principais conjuntos de dados públicos de Retinopatia Diabética disponíveis na literatura para detecção e segmentação de lesões de fundo. Foram discutidas características destes conjuntos de dados de imagens de fundo, como a quantidade e qualidade das imagens disponíveis nestes conjuntos de dados, a disponibilização das anotações em nível de *pixel* das lesões da retina, e a disponibilização de *Ground Truth* realizado por múltiplos especialistas. Por fim, foi apresentada uma tabela comparativa que sumariza as principais características dos conjuntos de dados públicos de imagens de fundo investigados.

No próximo Capítulo serão apresentados e discutidas as principais características e problemas de visão computacional, assim como conceitos básicos sobre aumento de dados, transferência de aprendizado, métricas e modelos para detecção e segmentação de objetos.

[12] https://ieee-dataport.org/open-access/indian-diabetic-retinopathy-image-dataset-idrid

4 APRENDIZADO PROFUNDO NA DETECÇÃO DE OBJETOS

Este Capítulo tem por objetivo apresentar os conceitos básicos sobre aprendizado profundo na detecção de objetos. São apresentados modelos de redes neurais convolucionais profundas no estado da arte, assim como são discutidas as técnicas de aumento de dados e transferência de aprendizado. Por fim, são apresentadas diferentes métricas de desempenho para avaliar os modelos preditivos.

4.1 Introdução

Redes neurais profundas têm sido aplicadas com sucesso em tarefas de Visão Computacional (VC), mais especificamente na classificação, detecção e segmentação de imagens digitais, em função da evolução das Redes Neurais Convolucionais (em inglês, *Convolutional Neural Network* – CNN) (GU et al., 2017; ALOYSIUS; GEETHA, 2017; SHORTEN; KHOSHGOFTAAR, 2019; JIAO et al., 2019; ZOU et al., 2019; ZHAO et al., 2019; MURTHY et al., 2020). Assim, antes de abordar o funcionamento de algoritmos de detecção de objetos é necessário compreender o propósito das CNNs, que são os blocos de construção básicos para a maioria dos modelos de aprendizado profundo.

Rede Neural Convolucional é uma espécie de rede neural *feedforward*, em que a saída de uma camada é usada como entrada para a próxima camada, sendo capaz de extrair características de dados por meio de convoluções (LI et al., 2020). Diferente dos métodos tradicionais de extração de características (LINDEBERG, 2012; DALAL; TRIGGS, 2005; AHONEN; HADID; PIETIKAINEN, 2006), a CNN não precisa extrair características manualmente. A arquitetura da CNN é inspirada pela percepção visual (LI et al., 2020). Krizhevsky; Sutskever; Hinton (2012) propôs uma arquitetura clássica de CNN que apresentou melhorias em relação aos métodos anteriores que eram utilizados na tarefa de classificação de imagens (ZHANG et al., 2019).

Essas redes neurais utilizam filtros (*kernels*) parametrizados e esparsamente conectados que preservam as características espaciais das imagens. As camadas con-

volucionais reduzem sequencialmente a resolução espacial das imagens enquanto expandem a profundidade de seus mapas de características. Essa série de transformações convolucionais pode criar representações de imagens em dimensões muito mais baixas e mais úteis do que se fossem realizadas manualmente (SHORTEN; KHOSHGOFTAAR, 2019).

De acordo com Shiloh-perl; Giryes (2020); Li; Krishna; Xu (2021), atualmente as CNNs têm sido amplamente utilizadas para resolver problemas de VC, tais como: (1) classificação (YUDITA; MANTORO; AYU, 2021; FUAD et al., 2021), como o reconhecimento de câncer em imagens médicas (JAKIMOVSKI; DAVCEV, 2018; LI et al., 2020; CHORIANOPOULOS et al., 2020); (2) classificação e localização, em que o objetivo não é apenas saber se uma imagem contém um determinado objeto, mas também onde exatamente está este objeto na imagem; (3) detecção e localização de vários objetos, como na tarefa de detecção de objetos realizados por carros autônomos (KULKARNI; DHAVALIKAR; BANGAR, 2018; AKYOL et al., 2020); (4) segmentação semântica, que envolve a detecção do conjunto de *pixels* pertencentes a uma classe específica de um objeto (JEONG et al., 2020; PHAM, 2021), cujo objetivo é identificar o objeto alvo com maior proximidade, atribuindo cada *pixel* a uma determinada classe (NELSON, 2020); e, (5) segmentação de instância, em que há uma diferenciação entre objetos de uma mesma classe (POTLAPALLY et al., 2019; TIAN; YUAN; LIU, 2020; KUMAR; ZHANG, 2021), sendo possível diferenciar entre classes de objetos, e objetos dentro de cada classe.

4.2 Modelos para Detecção

A detecção de objetos é um problema de Visão Computacional que tem por objetivo identificar e localizar objetos pertencentes a uma determinada classe em uma imagem (GANESH, 2019). A interpretação da localização do objeto pode ser realizada de diferentes formas, incluindo a criação de uma caixa delimitadora em torno do objeto detectado, ou com a marcação de cada *pixel* na imagem que contém o objeto (segmentação).

O objetivo de detectar objetos está associado à capacidade de identificar, diferenciar e classificar objetos rapidamente, permitindo a tomada de decisões rápidas em relação aos objetos detectados. A seguir, serão apresentados modelos de aprendizado profundo no estado da arte capazes de realizar a detecção de objetos em imagens digitais.

4.2.1 *Regions with Convolutional Neural Network features*

Este método, proposto por Girshick et al. (2014), tem o propósito de combinar propostas de região com redes neurais convolucionais (em inglês, *Regions with Con-*

volutional Neural Network features – R-CNN) (ZHAO et al., 2019). É uma das abordagens de detecção de objetos de aprendizagem profunda que funciona de acordo com duas ideias principais: (1) aplicação de redes neurais convolucionais de alta capacidade a propostas de região, de baixo para cima, para localizar e segmentar objetos; e, (2) quando os dados de treinamento rotulados são escassos, a realização de pré-treinamento supervisionado como tarefa auxiliar, seguida por um ajuste fino no domínio do problema (GIRSHICK et al., 2014).

É um algoritmo de detecção de dois estágios (LI; KRISHNA; XU, 2021), sendo que o primeiro estágio identifica um subconjunto de regiões em uma imagem que pode conter um objeto e o segundo estágio classifica o objeto em cada região. O modelo realiza as três etapas a seguir:

1. Encontra as regiões na imagem que podem conter um objeto. Essas regiões são chamadas de propostas de região.
2. Extrai as características das regiões selecionadas utilizando CNN.
3. Classifica os objetos utilizando as características extraídas.

A Figura 11 apresenta uma visão geral do sistema de detecção de objetos do modelo R-CNN, em que: (1) o modelo lê uma imagem de entrada; (2) extrai cerca de 2.000 propostas de regiões; (3) calcula as características para cada proposta usando uma CNN; e, então, (4) classifica cada região usando Máquinas de Vetores de Suporte (em inglês, *Support Vector Machines* – SVM) (LI; KRISHNA; XU, 2021).

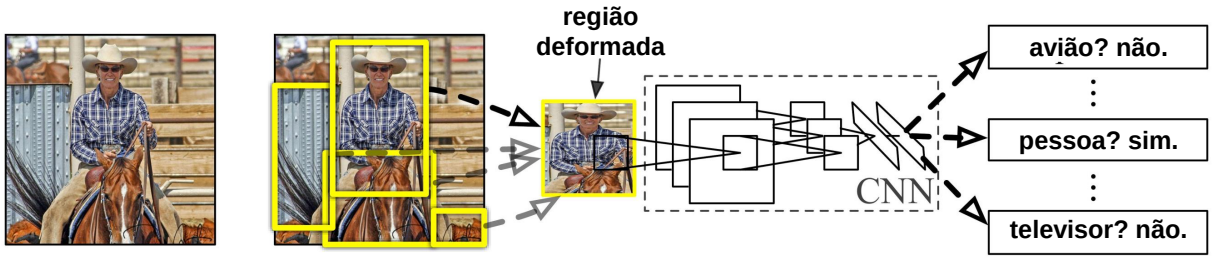


Figura 11 – Visão Geral do Sistema de Detecção de Objetos do modelo R-CNN. Fonte: Adaptado de Girshick et al. (2014).

O modelo R-CNN produziu um aumento significativo de desempenho em relação aos métodos que o precederam, entretanto seu custo computacional é elevado, uma vez que a CNN utilizada como extrator de características é executada para cada proposta de região separadamente (GU et al., 2017). A seguir são apresentados modelos baseados em propostas de regiões que apresentaram inovações e desempenho superior à R-CNN (ZHAO et al., 2019; LI; KRISHNA; XU, 2021).

4.2.2 *Fast Regions with Convolutional Neural Network features*

Assim como o modelo R-CNN, o modelo Fast R-CNN (em inglês, *Fast Regions with Convolutional Neural Network features*) (GIRSHICK, 2015), também gera propostas de região. Porém, ao contrário do detector R-CNN, que recorta e redimensiona as propostas de região, o detector Fast R-CNN processa a imagem inteira.

Enquanto o detector R-CNN classifica cada região, o modelo Fast R-CNN agrupa características correspondentes da CNN a cada proposta de região. O Fast R-CNN é mais eficiente que o R-CNN porque os cálculos para regiões sobrepostas são compartilhados (ZHAO et al., 2019; LI; KRISHNA; XU, 2021).

O modelo Fast R-CNN carrega toda a imagem e as propostas de região como entrada em sua arquitetura CNN em uma propagação direta. Além disso, combina diferentes partes da arquitetura, tais como CNNs, Regiões de Interesse (em inglês, *Region of Interest* – ROI) e a camada de classificação em uma arquitetura. Este modelo também utiliza uma camada *Softmax* ao invés de SVM para classificar a proposta de região, tornando o modelo mais rápido em comparação ao modelo R-CNN (GIRSHICK, 2015).

A Figura 12 apresenta a arquitetura do modelo Fast R-CNN, em que uma imagem de entrada e várias ROIs são inseridas em uma rede neural convolucional. Cada ROI é agrupada em um mapa de características de tamanho fixo e em seguida, mapeada para um vetor de características por camadas totalmente conectadas (FC). A rede tem dois vetores de saída por ROI: camada *Softmax* (que classifica os objetos) e camada de regressão de caixa delimitadora (que localiza os objetos). A arquitetura é treinada de ponta a ponta com uma função de perda multitarefa (GIRSHICK, 2015).

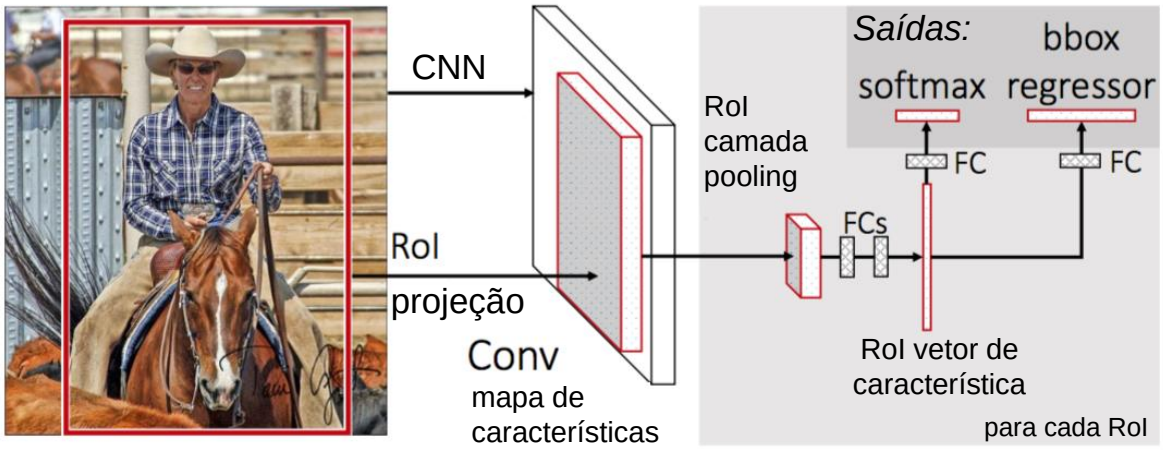


Figura 12 – Visão Geral do Sistema de Detecção de Objetos do modelo Fast R-CNN. Fonte: Adaptado de Girshick (2015).

De acordo com Gu et al. (2017), o modelo Fast R-CNN apresentou melhorias em relação ao R-CNN, como um método de treinamento de ponta a ponta, em que todas

as camadas de rede podem ser atualizadas durante o ajuste fino, que simplifica o processo de aprendizagem e melhora a precisão da detecção dos objetos. Segundo Li; Krishna; Xu (2021), o modelo Fast R-CNN apresenta as seguintes vantagens:

- Mais rápido que o modelo R-CNN porque não é preciso carregar 2.000 propostas de região na CNN em todas as situações.
- A operação de convolução é realizada apenas uma vez por imagem e um mapa de características é gerado a partir desta convolução.

Entretanto, o modelo Fast R-CNN possui algumas limitações, dentre as quais: a necessidade de um algoritmo de seleção de regiões; e, o tempo de inferência elevado (SANTOS, 2018). Embora o modelo Fast R-CNN seja mais eficiente que o modelo R-CNN, sua velocidade de detecção ainda é limitada pela detecção das propostas de regiões (ZOU et al., 2019). Nesse contexto, a seguir é apresentado o modelo Faster R-CNN, que apresenta inovações que visam contornar as limitações observadas no modelo Fast R-CNN.

4.2.3 *Faster Regions with Convolutional Neural Network features*

O modelo Faster R-CNN (em inglês, *Faster Regions with Convolutional Neural Network features*) é composto por dois módulos. O primeiro módulo é uma rede neural profunda totalmente convolucional que propõe as regiões e o segundo módulo é o detector Fast R-CNN (REN et al., 2017). O principal obstáculo da arquitetura do Fast R-CNN, lançado em 2015, continuava sendo o algoritmo de busca seletiva, pois era necessário gerar 2.000 propostas por imagem, impactando no tempo de treinamento do modelo.

No modelo Faster R-CNN, a busca seletiva foi substituída pela Rede de Proposta de Região (em inglês, *Region Proposal Networks* – RPN) (REN et al., 2017; ZHAO et al., 2019). Primeiramente, a imagem é passada para o *Backbone*), que por sua vez gera um mapa de características da convolução. Esses mapas de características então são passados para a Rede de Proposta de Região. Então, a RPN pega um mapa de características e gera as âncoras (centro da janela deslizante com um tamanho e escala únicos). Estas âncoras são então passadas para a camada de classificação (que classifica o objeto em uma determinada classe) e a camada de regressão (que localiza a caixa delimitadora associada a um objeto). A Figura 13 ilustra o módulo RPN introduzido na Faster R-CNN (REN et al., 2017).

A principal contribuição do modelo Faster R-CNN foi a introdução da RPN para geração de propostas de objetos (ZHAO et al., 2019; ZOU et al., 2019). Com as melhorias apresentadas na arquitetura do Faster R-CNN o modelo tornou-se mais rápido e mais preciso que os modelos de mesmo propósito que o antecederam (GU et al.,

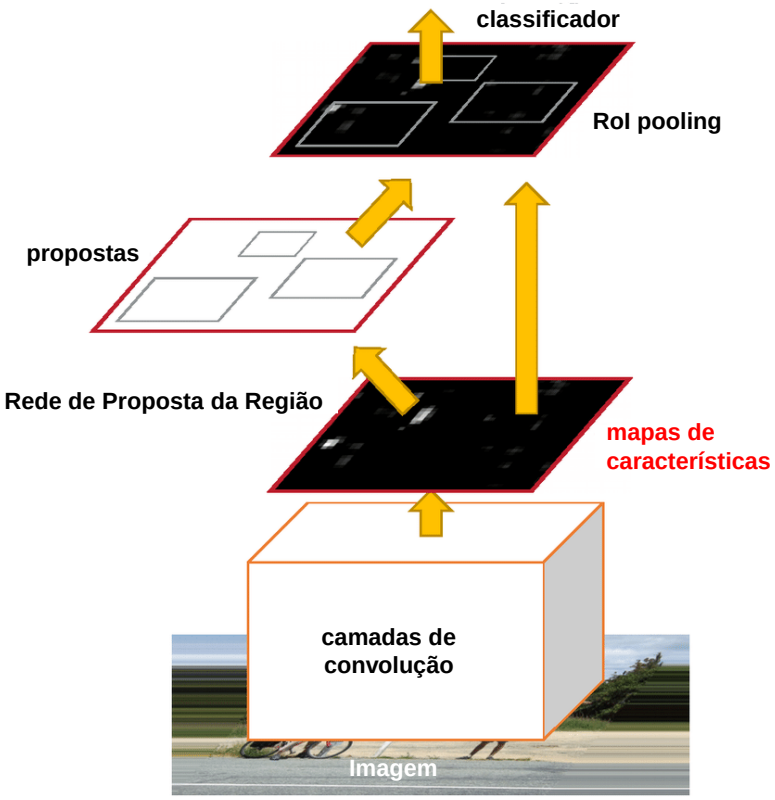


Figura 13 – Módulo RPN introduzido no modelo Faster R-CNN. Fonte: Adaptado de Ren et al. (2017).

2017). Embora o modelo Faster R-CNN seja mais rápido que o modelo Fast R-CNN, ainda há redundância de computação na fase de detecção (ZOU et al., 2019), que torna o modelo pouco eficiente para utilização em sistemas embarcados ou de tempo real. Ainda assim, o modelo Faster R-CNN é referência na detecção de objetos, servindo de base para o desenvolvimento de diversos outros modelos (SANTOS, 2018).

4.2.4 *You Only Look Once*

Proposto por Redmon et al. (2016), o modelo YOLO (em inglês, *You Only Look Once*) é uma abordagem mais rápida para a detecção de objetos em comparação às R-CNNs discutidas anteriormente (MURTHY et al., 2020). Basicamente, modelos baseados em propostas de região usam um classificador para detectar o objeto e, em seguida, classificam sua presença em vários locais (regiões) da imagem. Nesse sentido, estes modelos tornam-se lentos e de difícil otimização.

De acordo com Zou et al. (2019) e Murthy et al. (2020), YOLO foi o primeiro detector de estágio único na era do aprendizado profundo e descontinuou o paradigma adotado pelos detectores anteriores, baseados em propostas de região, seguindo uma abordagem totalmente diferente, em que é aplicada uma única rede neural na imagem completa, dividindo a imagem em grades e prevendo as caixas delimitadoras e probabilidades para cada grade simultaneamente.

O YOLO surgiu como alternativa pois considera a detecção de objetos como um problema de regressão, cujo propósito é identificar as caixas delimitadoras espacialmente separadas e as probabilidades de ocorrência das classes associadas. Uma única rede neural prevê as caixas delimitadoras e as probabilidades de ocorrência de classe diretamente na imagem completa que está sendo avaliada. Como todo o *pipeline* de detecção é uma única rede, este processo pode ser otimizado de ponta a ponta, melhorando o desempenho da detecção (REDMON et al., 2016).

Os modelos de detecção de objetos que se baseiam em regiões para localizar um objeto na imagem não olham a imagem completa mas partes da imagem que têm maior probabilidade de conter o objeto. Em contrapartida, o YOLO é um modelo de detecção de objetos que utiliza uma única rede convolucional para prever as caixas delimitadoras e as probabilidades de classe para essas caixas (ZHAO et al., 2019).

Para realizar a detecção de objetos utilizando o modelo YOLO, inicialmente uma imagem é dividida em uma grade com células S$\times$S (MURTHY et al., 2020). Em seguida, dentro de cada grade são selecionadas *m* caixas delimitadoras. Para cada caixa, a rede gera uma probabilidade de classe e valores de deslocamento para esta caixa. As caixas delimitadoras com probabilidade de classe acima de um valor limite são selecionadas e usadas para localizar o objeto na imagem (REDMON et al., 2016; JIAO et al., 2019).

A Figura 14 apresenta o funcionamento do modelo YOLO, em que a detecção é resolvida como um problema de regressão. A imagem é divida em uma grade S$\times$S e, para cada célula da grade, são previstas as caixas delimitadoras, a confiança para cada uma dessas caixas e as probabilidades de ocorrência de cada classe (ZHAO et al., 2019).

A seguir são apresentadas algumas vantagens e desvantagens do modelo YOLO (REDMON et al., 2016; SANTOS, 2018; ZOU et al., 2019; JIAO et al., 2019; ZHAO et al., 2019):

- Em função de sua detecção ser realizada em estágio único, é mais rápido que arquiteturas que realizam a detecção em dois estágios.
- As versões mais atuais apresentam desempenho superior na detecção de objetos em diferentes escalas.
- A principal limitação reside na dificuldade em detectar objetos pequenos, principalmente quando estes aparecem agrupados na imagem (MURTHY et al., 2020).
- Restrições espaciais em função do modelo possuir um limite de previsões de caixas limitadoras para cada célula.

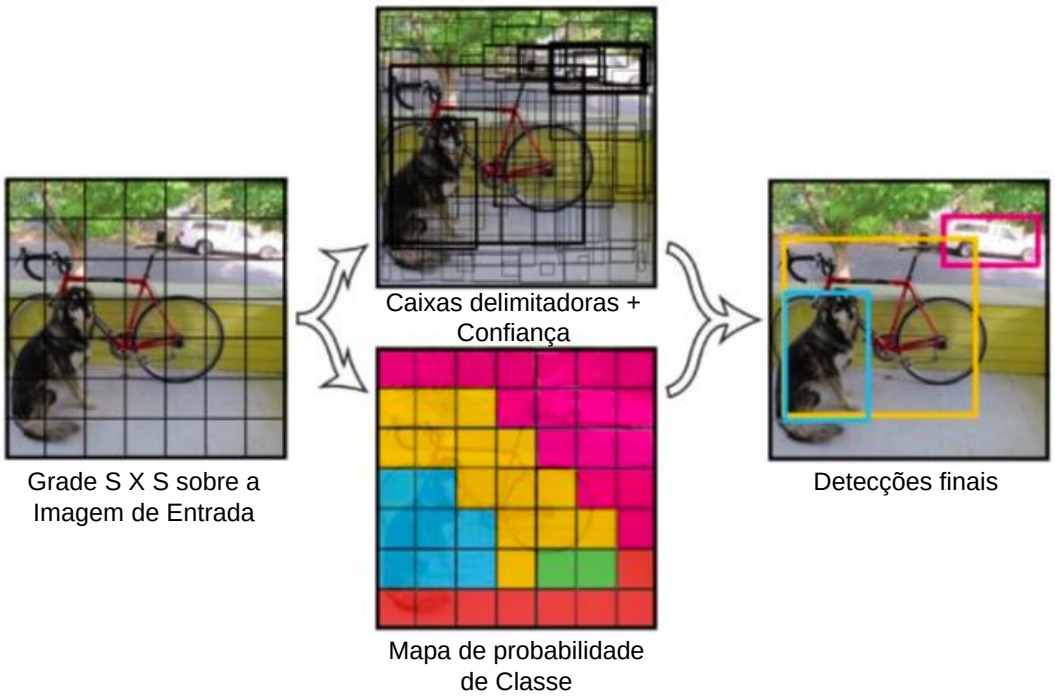


Figura 14 – Funcionamento do modelo YOLO. Fonte: Adaptado de Redmon et al. (2016).

Desde sua primeira versão, uma série de melhorias foram introduzidas no YOLO. Algumas destas melhorias podem ser verificadas na versão 2 (REDMON; FARHADI, 2017), versão 3 (REDMON; FARHADI, 2018; ZHENG; ZHAO; LI, 2021; HOSSAIN et al., 2021), versão 4 (BOCHKOVSKIY; WANG; LIAO, 2020; SANTOS et al., 2021) e versão 5 (XU et al., 2021; ZHU et al., 2021).

Em suma, os detectores de objetos em imagens digitais podem ser divididos basicamente em duas categorias: a primeira, composta pelos detectores de dois estágios, sendo o modelo Faster R-CNN (REN et al., 2017) o mais representativo desta categoria; e, a segunda, composta por detectores de estágio único, como por exemplo o modelo YOLO (REDMON et al., 2016). No geral, os detectores de dois estágios costumam apresentar maior precisão na localização e classificação de objetos, enquanto que os detectores de estágio único costumam ser mais rápidos (JIAO et al., 2019; KIM; SUNG; PARK, 2020).

4.3 Modelo para Segmentação de Instância

Em contraste com a segmentação clássica de imagens baseada em visão computacional por meio do valor da escala de cinza (valor de *pixel*), a segmentação de imagens baseada em redes neurais pode ser categorizada em três tipos principais: segmentação semântica, segmentação de instância e segmentação panóptica. A segmentação de imagens envolve a classificação de cada *pixel* em uma imagem com o rótulo correto, de modo que os *pixels* que compartilham o mesmo rótulo tenham certas

características. A segmentação semântica é um problema de classificação em nível de *pixel* com rótulos semânticos (ou seja, um conjunto de objetos). A segmentação de instância realiza a classificação de *pixels* para o particionamento de objetos individuais. Na verdade, a segmentação de instância melhora o escopo e a capacidade da segmentação semântica ao adotar a detecção e retratar cada objeto de interesse da imagem (ZOU, 2023)

Portanto, na segmentação de instância é gerado um mapa de segmentação para cada instância detectada de um objeto. A segmentação de instância trata objetos individuais como entidades distintas, independentemente da classe dos objetos, diferentemente da segmentação semântica, que considera todos os objetos da mesma classe como pertencentes a uma única entidade. A seguir, será apresentado o modelo Mask R-CNN, que se estende na arquitetura do modelo Faster R-CNN para realizar a segmentação em nível de *pixel* dos objetos detectados (HE et al., 2020).

4.3.1 *Mask Regions with Convolutional Neural Network features*

O modelo Mask R-CNN (em inglês, *Mask Regions with Convolutional Neural Network features*), consiste de dois estágios. O primeiro estágio é composto por uma rede de proposta de região, que prevê as caixas delimitadoras de propostas de objetos com base em âncoras (conjunto de caixas com localizações predefinidas). O segundo estágio é composto por um detector R-CNN que refina essas propostas, classificando-as e calculando a segmentação em nível de *pixel* para essas propostas (HE et al., 2020).

De acordo com Arnab; Torr (2017), a segmentação de instância é uma tarefa que requer a detecção de todos os objetos em uma imagem e a segmentação semântica (DAI; HE; SUN, 2015) de cada instância. As tarefas de detecção e segmentação são geralmente consideradas como dois processos independentes, porém, quando realizadas separadamente para realizar a segmentação de instância, podem gerar bordas espúrias e sobreposição de instâncias (LI et al., 2017).

Para resolver este problema, o modelo Mask R-CNN utiliza as ramificações de classificação e regressão de caixa delimitadora do modelo Faster R-CNN, adicionando paralelamente uma ramificação para prever a segmentação das máscaras *pixel* a *pixel* (ZHAO et al., 2019). Modelos de detecção de objetos como YOLO e R-CNN traçam uma caixa delimitadora ao redor dos objetos detectados, enquanto que na segmentação de instância são fornecidas as máscaras em *pixels* para cada objeto detectado na imagem.

Mask R-CNN estende o modelo Faster R-CNN (ZHAO et al., 2019; JIAO et al., 2019; MELO et al., 2020), adicionando uma ramificação de máscara para segmentação de objetos em nível de *pixel*. Baseia-se no conceito de regiões de interesse, em que uma *Feature Pyramid Network* é utilizada para gerar os mapas de características

das imagens. Em seguida, uma RPN analisa os mapas de características gerados pela FPN para encontrar regiões da imagem (âncoras) capazes de representar um objeto (RAMCHARAN et al., 2018; REN et al., 2017; OJHA; SAHU; DEWANGAN, 2021).

A Figura 15 apresenta a estrutura do modelo Mask R-CNN para realizar segmentação de instância, em que é possível observar que o modelo baseia-se na estrutura do Faster R-CNN, sendo adicionada uma terceira ramificação para prever a máscara do objeto detectado em paralelo com as ramificações existentes para classificação e localização.

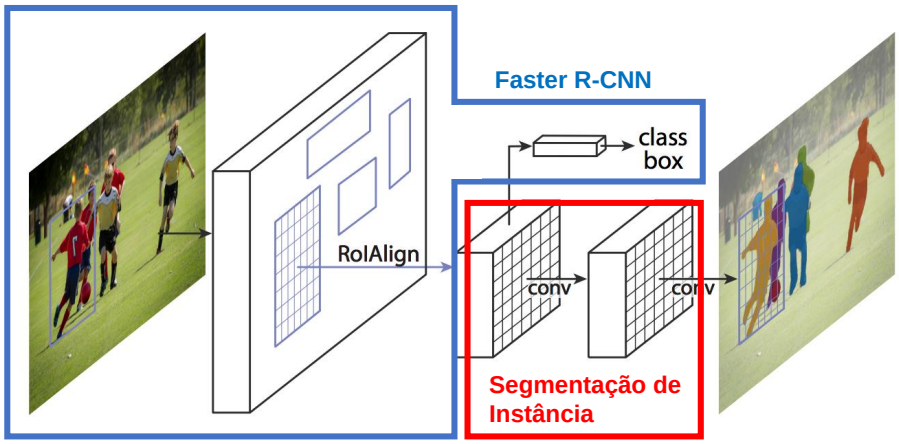


Figura 15 – Estrutura do Mask R-CNN para Segmentação de Instância. Fonte: Adaptado de He et al. (2020).

As âncoras podem formar regiões de interesse de vários tamanhos, então a operação ROIAlign (em inglês, *Region of Interest Align*)[1] (HE et al., 2020), redimensiona estas regiões para um único tamanho. Na sequência, as camadas convolucionais e totalmente conectadas realizam a predição. O Mask R-CNN tem três saídas: a classe, a caixa delimitadora e a máscara de segmentação (JIAO et al., 2019; GONZALEZ; ARELLANO; TAPIA, 2019; HE et al., 2020).

A inserção de uma ramificação de máscara na estrutura do modelo Faster R-CNN adiciona uma pequena carga computacional, porém sua combinação com as tarefas de classificação e regressão de caixa delimitadora fornece informações complementares para o objeto detectado, fazendo com que o modelo Mask R-CNN torne-se uma estrutura flexível e eficiente (JIAO et al., 2019), sendo portanto um bom candidato para tarefas de detecção em nível de segmentação de instância (ZHAO et al., 2019).

[1]Operação realizada para extrair um pequeno mapa de características de cada ROI em tarefas baseadas em detecção e segmentação.

4.4 Aumento de Dados

Abordagens baseadas em aprendizado profundo necessitam de grandes volumes de dados de amostras para treinamento. Desse modo, se o conjunto de dados possui uma quantidade limitada de imagens de amostra, por exemplo, pode ser necessário criar novas amostras que são adicionadas aos dados disponíveis. Esta técnica é conhecida como aumento de dados (GOODFELLOW et al., 2020).

De maneira geral, problemas que envolvem análise de imagens da retina, mais especificamente a detecção das lesões de fundo, contam com conjuntos de dados relativamente pequenos. O baixo número de amostras destes conjuntos de dados pode ocasionar subajuste (*underfitting*) durante o treinamento dos modelos responsáveis pela detecção das lesões de fundo. Nestes casos, os modelos não conseguem extrair informações suficientes sobre o conjunto de dados, levando a taxas elevadas de erros. Conjuntos de dados maiores resultam em modelos de Aprendizado Profundo com melhor capacidade de predição (HALEVY; NORVIG; PEREIRA, 2009; SUN et al., 2017). No entanto, reunir enormes conjuntos de dados pode ser uma tarefa difícil devido ao esforço manual de coletar e rotular os dados.

Construir grandes conjuntos de dados de imagens médicas envolve muitos desafios, tais como: a raridade das doenças; o direito de privacidade dos pacientes; a exigência de médicos especialistas para a realização das anotações das doenças; e, os custos e esforços manuais necessários para conduzir a aquisição e processamento destas imagens.

O aumento de dados é uma técnica eficaz para se obter conjuntos de dados com quantidades maiores de amostras, representando um conjunto mais abrangente e significativo de dados de amostra, minimizando o problema de escassez de informações para treinamento de redes neurais profundas e potencializando a capacidade de predição destes modelos.

O aumento de dados pode melhorar o processo de aprendizagem e capacidade de generalização de um modelo (KRIZHEVSKY; SUTSKEVER; HINTON, 2012; ZEILER; FERGUS, 2014) e pode ser realizado por transformações sobre as imagens originais, tais como (PIRES; ROCHA; WAINER, 2019): (1) transformações geométricas, como escala, translações, rotações, recorte, etc.; ou, (2) transformações fotométricas, como equalizações de histograma, aprimoramentos de contraste, etc. A Figura 16, apresenta uma ilustração com imagens artificiais criadas por meio de técnicas de aumento de dados a partir de uma imagem original de ressonância magnética para segmentação de tumor cerebral (NALEPA; MARCINKIEWICZ; KAWULOK, 2019).

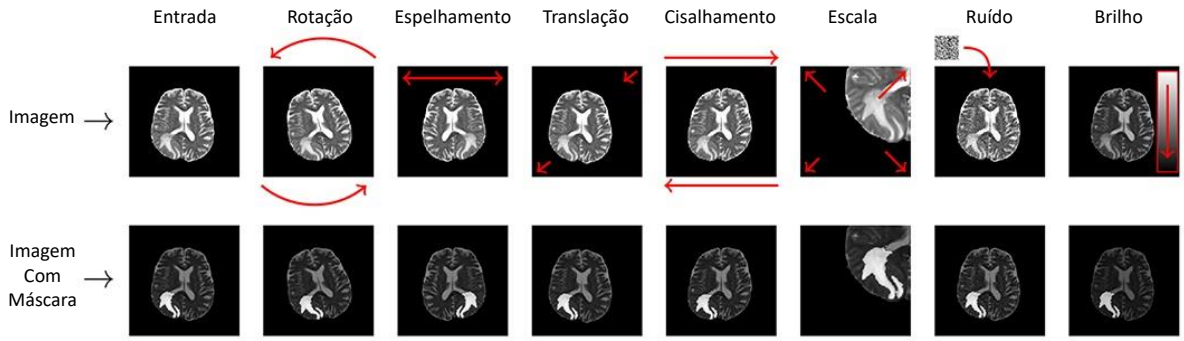


Figura 16 – Exemplo de aumento de dados com a geração de sete novas imagens com base em uma imagem original de ressonância magnética juntamente com sua máscara (*Ground Truth*). Fonte: Adaptado de Nalepa; Marcinkiewicz; Kawulok (2019).

O aumento de imagens cria novos exemplos de treinamento a partir de dados de treinamento já existentes. É possível gerar artificialmente diferentes imagens para cada imagem original, possibilizando um treinamento mais eficaz e permitindo que o modelo seja capaz de realizar inferências para uma quantidade maior situações.

4.5 Transferência de Aprendizado

A transferência de aprendizado tem por objetivo transferir o conhecimento aprendido em uma ou mais tarefas de origem para melhorar o processo de aprendizagem em uma tarefa alvo, buscando evitar o *overfitting* (PIRES; ROCHA; WAINER, 2019; WEISS; KHOSHGOFTAAR; WANG, 2016; SHAO; ZHU; LI, 2015). Por exemplo, uma rede pode ser treinada em um grande conjunto de dados como o ImageNet (FEI-FEI; DENG; LI, 2010) e após usam-se os pesos treinados desta rede como os pesos iniciais em uma nova tarefa de classificação (SHORTEN; KHOSHGOFTAAR, 2019).

Segundo Bishop; Coombs; Henry (1973); Andrews; Pollen (1979); Jones; Palmer (1987); Sabatini (1996), o perfil dos campos receptivos de células simples no córtex dos mamíferos pode ser descrito por funções Gabor bidimensionais orientadas e estes filtros Gabor são usados no processamento de imagens para extração de características e análise de textura (AACH; KAUP; MESTER, 1995). Yosinski et al. (2014) descobrem que na primeira camada são aprendidas características semelhantes aos filtros Gabor (AL-KADI, 2017). Essas características obtidas nas primeiras camadas não são específicas de um conjunto de dados e podem ser generalizadas para tarefas que envolvam a classificação de imagens, mesmo que em domínios diferentes.

De acordo com Vrbancic; Podgorelec (2020), a utilização de transferência de aprendizado geralmente é eficaz, pois muitos conjuntos de dados de imagens compartilham características espaciais de baixo nível que são melhor aprendidas com grande volumes de dados. Zamir et al. (2019) mostram que o número total de dados rotulados necessários para resolver um conjunto de tarefas pode ser reduzido em cerca de

2/3 (em comparação com o treinamento sem a aplicação da transferência de aprendizado), mantendo praticamente a mesma precisão e diminuindo o custo computacional necessário para o treinamento do modelo.

Esta técnica normalmente é utilizada quando não há amostras de treinamento suficientes no domínio da tarefa alvo, ou quando já existe um solução razoável para um problema relacionado. Nestas situações, pode ser mais rápido reutilizar pesos de uma rede pré-treinada para resolver o problema-alvo. Uma parte dos trabalhos existentes (PAN; YANG, 2010), que exploram transferência de aprendizado, pressupõe que os domínios de origem e destino devam ter alguma relação entre si, embora esta técnica tenha apresentado bons resultados mesmo quando aplicada em domínios diferentes (VRBANCIC; PODGORELEC, 2020; HORRY et al., 2020; SANTOS et al., 2021).

A transferência de aprendizado frequentemente aparece em dois cenários: na extração de características e no ajuste fino (em inglês, *fine tuning*) do modelo. No primeiro cenário, são congeladas as camadas de rede e os parâmetros correspondentes para utilizá-los apenas para extrair características para a tarefa de destino. Já no segundo, envolve-se a continuação da retro-propagação dos pesos e atualização dos parâmetros internos da rede a fim de ajustá-los ao domínio de destino (REYES; CAICEDO; CAMARGO, 2015; VRBANCIC; PODGORELEC, 2020).

4.6 Métricas de Desempenho

As métricas adotadas para avaliar um modelo devem estar de acordo com o tipo de tarefa que será avaliada, sendo importante definir quais são as métricas mais apropriadas para aferir o desempenho do modelo. Em tarefas de detecção, a avaliação quantitativa é realizada pela estimativa de sobreposição de regiões entre as imagens detectadas e as anotações de caixas delimitadoras dos objetos nas imagens originais (*Ground Truth*).

A partir dessa comparação, é possível obter quatro classes de classificação (SILVA, 2019): Verdadeiros Positivos (VP), representando uma classificação correta da classe Positiva; Falsos Positivos (FP), representando uma classificação errada para a classe Positiva, quando o resultado real era para ser da classe Negativa; Verdadeiros Negativos (VN), representando uma classificação correta da classe Negativa; e, Falsos Negativos (FN), representando uma classificação errada para a classe Negativa, quando o resultado real era para ser da classe Positiva.

Com as informações descritas acima, é possível realizar o cálculo de algumas métricas associadas à classificação, tais como a Acurácia (em inglês, *Accuracy*), Precisão (HSU; LIN, 2021; DEWI et al., 2021), Revocação (HSU; LIN, 2021; DEWI et al., 2021) e F1-*score* (MOHAMMADIAN; KARSAZ; ROSHAN, 2017; HAQUE; AB-

DUR RAHMAN; SIDDIK, 2019; POWERS, 2020; DEWI et al., 2021).

A Acurácia tem a função de dar uma visão geral do desempenho do modelo, mostrando quanto este classificou corretamente. A Precisão se encarrega de avaliar as classificações da classe Positiva, dando um panorama de quantas estão corretas. A Revocação, por sua vez, pega o apanhado das classes Positivas como valor esperado, e quantifica quantas estão corretas. O F1-*Score* elucida a média harmônica entre a Precisão e a Revocação (WANGENHEIM, 2020).

No entanto, a métrica padrão para avaliar a precisão em problemas de classificação de imagens não pode ser aplicada diretamente em problemas de detecção ou segmentação de objetos, pois neste tipo de problema cada imagem pode ter objetos diferentes de classes diferentes e tanto a classificação quanto a localização dos objetos precisam ser avaliadas. Para modelos que realizam a detecção ou a segmentação de instância de objetos em imagens, as métricas são avaliadas com base no *Ground Truth* presente nas imagens de treinamento. O *Ground Truth* inclui a imagem, as classes dos objetos presentes na imagem e as anotações de caixas delimitadoras (em inglês, *bounding boxes*) de cada um dos objetos contidos na imagem.

Portanto, as métricas que avaliam estes modelos têm por objetivo comparar o que foi anotado na imagem e o que foi predito. Como exemplos de métricas frequentemente utilizadas na literatura para este tipo de problema, citam-se: a Interseção sobre a União (em inglês, *Intersection Over Union* – IoU) (REZATOFIGHI et al., 2019; MELO et al., 2020); a Precisão Média (em inglês, *Average Precision* – AP) (BOCHKOVSKIY; WANG; LIAO, 2020; HE et al., 2020; MAMDOUH; KHATTAB, 2021; DEWI et al., 2021); e, a média da Precisão Média (em inglês, *mean Average Precision* – mAP) (REDMON et al., 2016; REN et al., 2017; REDMON; FARHADI, 2018; MELO et al., 2020) (HOSANG et al., 2016; AIDOUNI, 2019).

O IoU, também identificado por coeficiente de similaridade de Jaccard é a métrica de avaliação mais popular usada nos *benchmarks* de detecção de objetos, sendo uma estatística para estimar a similaridade entre dois conjuntos de amostras (IACOVACCI; WU; BIANCONI, 2015; BERTELS et al., 2019; REZATOFIGHI et al., 2019). Mede a similaridade entre dois conjuntos finitos e é dada pela Equação 1 (WANGENHEIM, 2020; MAMDOUH; KHATTAB, 2021):

$$IoU = \frac{Area\ Overlap}{Area\ Union} = \frac{area(B_p \cap B_{gt})}{area(B_p \cup B_{gt})} = \frac{\text{VP}}{\text{VP} + \text{FP} + \text{FN}} \tag{1}$$

onde B_p, se refere à caixa com a predição do objeto (em inglês, *Box Predict* – B_p) e o B_{gt} se refere à caixa com a anotação original do objeto predito (em inglês, *Box Ground Truth* – B_{gt}).

Cada objeto é previsto por uma caixa delimitadora e quando uma predição é realizada, gerando-se várias detecções para um mesmo objeto, pode ocorrer uma sobreposição de caixas delimitadoras para um mesmo objeto, fazendo-se necessária a aplicação da técnica de Supressão de não Máximos (em inglês, *Non-max Suppression* – NMS) (BODLA et al., 2017). Esta técnica considera apenas a caixa delimitadora com o IoU mais alto, de forma que as demais caixas delimitadoras previstas para um mesmo objeto são descartadas (MAMDOUH; KHATTAB, 2021).

A Área de Sobreposição (em inglês, *Area Overlap*) é obtida pela sobreposição entre a caixa delimitadora predita e a caixa delimitadora verdade (anotação original); e, a Área de União (em inglês, *Area Union*) é a área abrangida tanto pela caixa delimitadora predita quanto pela caixa delimitadora verdade. Logo, a Interseção sobre União é obtida pela razão entre a Área de Sobreposição e a Área de União das caixas delimitadoras previstas na detecção dos objetos.

Para utilizar o IoU como métrica de avaliação, um limite de precisão deve ser adotado, como exemplo, um IoU com limiar de 0,5. Quanto maior o coeficiente em relação a este limiar melhor será a qualidade e precisão do modelo avaliado. No entanto, a escolha arbitrária de um limiar de IoU pode não refletir totalmente o desempenho de diferentes métodos, pois qualquer precisão de localização superior ao limite é tratada igualmente. Nestes casos, é possível calcular a precisão com base em vários limiares de IoU (REZATOFIGHI et al., 2019).

A Precisão Média (em inglês, *Average Precision* – AP) é a Área sob a Curva (em inglês, *Area Under the Curve* – AUC) de *Precision* $\times$ *Recall*, também chamada de curva PR (FREITAS, 2019). De acordo com Amorim (2020), para compreensão do cálculo da métrica AP no contexto da detecção de objetos, necessita-se realizar algumas considerações importantes, as quais seguem:

VP: é contabilizado como VP quando o objeto está presente no *Ground Truth* e o modelo foi capaz de detectá-lo. Normalmente é considerado como VP quando o IoU entre o objeto predito e o *Ground Truth* é superior a um limiar pré-determinado.

FP: quando o modelo não detecta um objeto que se encontra no *Ground Truth* em função deste objeto ter obtido um IoU inferior a um limiar pré-determinado, ou quando não há objetos na imagem e o modelo detectou algum.

VN: geralmente, na detecção de objetos todas as métricas que contêm VN são ignoradas (PADILLA; NETTO; SILVA, 2020), pois o objetivo é detectar objetos e não candidatos a não objetos, diferentemente de tarefas de classificação, nas quais é necessário decidir que cada instância seja considerada negativa ou positiva.

FN: quando o objeto encontra-se no *Ground truth* e o modelo não o detecta, ou seja quando o modelo não gera nenhuma predição para este objeto.

Quando há várias detecções para um mesmo objeto, a detecção com a IoU mais alta é considerada VP, enquanto as detecções restantes são consideradas FP. Se um objeto estiver presente na imagem e sua detecção tiver um limiar de IoU menor que o *Ground Truth*, a predição é considerada FP. Se o objeto não estiver na imagem, e o modelo o detecta, então a predição é considerada FP. Se o objeto encontra-se no *Ground truth* e o modelo não gera nenhuma predição para este objeto, a classe do objeto recebe FN. A Revocação e a Precisão são então calculadas para cada classe aplicando as fórmulas para cada imagem, conforme apresentado nas Equações 2 e 3, respectivamente:

$$\text{Revocação} = \frac{\text{VP}}{\text{VP} + \text{FN}} = \frac{\text{Objetos detectados corretamente}}{\text{Todos objetos do } \textit{Ground Truth}} \tag{2}$$

$$\text{Precisão} = \frac{\text{VP}}{\text{VP} + \text{FP}} = \frac{\text{Objetos detectados corretamente}}{\text{Todos objetos detectados}} \tag{3}$$

Precisão e Revocação são medidas úteis para aferir a eficiência de um modelo em predizer classes quando estas estão desbalanceadas. A curva PR apresenta a troca entre Precisão e Revocação para limiares diferentes, sendo que uma área grande sob a curva representa um modelo com alto valor de precisão (associado a uma baixa taxa de falsos positivos) e alto valor de revocação (associado a uma baixa taxa de falsos negativos).

Quando Precisão e Revocação têm valores elevados é um indicador que o modelo está sendo preciso em retornar a maioria dos resultados positivos. Um modelo com alto valor de Revocação, mas com baixo valor de Precisão retorna muitos resultados, mas muitas das classes preditas estão incorretas quando comparadas com as classes de treinamento. Ao passo que um modelo com alto valor de Precisão, mas baixo valor de Revocação é o oposto, retornando poucos resultados, mas muitas das classes preditas estão corretas quando comparadas com as classes de treinamento. A curva PR é uma importante ferramenta para analisar os resultados de um preditor.

A métrica de Precisão mede a precisão das predições, ou seja, é a porcentagem de predições corretas, enquanto que a métrica de Revocação é a capacidade de um modelo em encontrar todos os casos relevantes dentro de um conjunto de dados. Os valores de Revocação aumentam à medida que diminui os valores de FN. Já a Precisão tem um padrão em zigue-zague, em que desce com o aumento de FP, e sobe com o aumento de VP. Um modelo que realiza detecção de objetos pode ser considerado bom se sua precisão permanecer alta à medida que sua revocação aumenta (PADILLA; NETTO; SILVA, 2020).

Outra maneira de avaliar modelos que realizam a detecção de objetos é por meio da mAP, uma métrica amplamente utilizada para avaliar modelos de aprendizado profundo (REDMON et al., 2016; REN et al., 2017; REDMON; FARHADI, 2018; MELO

et al., 2020). Sua principal característica é a capacidade de comparar diferentes modelos, contrapondo a precisão com a revocação. A definição da métrica mAP para detecção de objetos foi formalizada pela primeira vez no desafio PASCAL VOC. Para calcular a mAP, realiza-se a média da *Average Precision* calculada para todas as classes de objetos (FREITAS, 2019), conforme apresentado na Equação 4.

$$mAP = \frac{1}{N} \sum_{i=1}^{N} AP_i \tag{4}$$

Embora não seja simples quantificar e interpretar os resultados de um modelo, mAP é uma métrica que auxilia no processo de avaliação de modelos de aprendizado profundo que realizam a detecção de objetos.

Dependendo da forma que estão distribuídos os exemplos nas classes do conjunto de dados de treinamento, os valores de Precisão Média podem variar de muito alto para algumas classes (geralmente a classe majoritária), a muito baixo para as classes com menos exemplos ou com ruídos nos dados. Portanto, o valor obtido com a métrica mAP nestes casos pode não ser um parâmetro confiável para avaliar o desempenho geral do modelo, sendo inclusive um bom indicador que é necessário realizar o balanceamento dos dados. Nestes casos, é aconselhável verificar as precisões médias das classes individualmente para avaliar o modelo.

O desafio PASCAL VOC usa mAP como uma métrica com um limite IoU de 0,5, enquanto no desafio COCO é calculada a média de mAP sobre diferentes limites IoU (0.5, 0.55, 0.6, 0.65, 0.7, 0.75, 0.8, 0.85, 0.9, 0.95), com um passo de 0.05. Por vezes, a métrica mAP é apresentada em artigos por mAP@[.05:.95]. Portanto, no desafio COCO a média do AP é calculada em todas as classes e também nos limites IoU definidos (HUI, 2018).

Importante observar que há situações, como no caso da competição COCO, que os termos AP e mAP são utilizados de forma intercambiável, pois nestes contextos as métricas AP e AR são calculadas sobre vários valores de IoU. Mais especificamente, são usados 10 limites IoU de 0,50:0,05:0,95, para calcular o AP e AR, não sendo portanto calculadas sobre um único IoU de 0,5 (HUI, 2018; AIDOUNI, 2019). Esta abordagem é adotada pois a média sobre IoUs recompensa os detectores com melhor capacidade de localização. Para estes casos, em que o AP é calculado sobre vários limites de IoU, é comum considerá-lo como sendo uma média da Precisão Média.

Ao avaliar AP em vários valores de limite IoU, o modelo penaliza a localização pobre e otimiza para uma boa localização porque a precisão da localização é avaliada com base no IoU entre o *Ground Truth* e a caixa delimitadora prevista, sendo a estratégia mais adequada para aplicações de detecção que requerem alta precisão, como em carros autônomos ou imagens médicas (AIDOUNI, 2019).

Por fim, convém destacar que diferentemente da detecção de objetos, no contexto da segmentação semântica os valores para VP, VN, FP e FN, e mais especificamente o IoU, as métricas são calculadas diretamente sobre os *pixels* das imagens, em que:

VP: é a quantidade de *pixels* classificados/segmentados na classe correta em relação ao *Ground Truth*.

FP: é a quantidade de *pixels* classificados/segmentados como uma classe igual a do VP onde deveria ser diferente.

VN: é a quantidade de *pixels* classificados/segmentados na classe correta em relação ao *Ground Truth* (que seja diferente da classe analisada em relação ao VP).

FN: é a quantidade de *pixels* classificados/segmentados como uma classe diferente do VP onde deveria ser igual.

4.7 Considerações sobre o Capítulo

Neste Capítulo, foram apresentadas e discutidas as principais características e desafios de problemas de visão computacional, como a classificação, a classificação e localização, a detecção e localização de vários objetos, a segmentação semântica, e a segmentação de instância.

Em seguida, foram apresentados e discutidos o funcionamento de modelos utilizados para a detecção de objetos em dois estágios, como a família de arquiteturas baseadas no modelo R-CNN, e detectores de estágio único, como o YOLO. Na sequência, foi discutido o modelo Mask R-CNN, uma arquitetura que se estende na Faster R-CNN, e que tem por propósito realizar a segmentação em nível de *pixel* dos objetos detectados.

Também foram apresentados conceitos básicos sobre técnicas de aumento de dados, transferência de aprendizado e diferentes métricas de desempenho para avaliar a precisão dos modelos discutidos. No próximo Capítulo serão abordados trabalhos relevantes que condizem com o estado da arte na identificação de lesões de fundo associadas à Retinopatia Diabética.

5 ESTADO DA ARTE

Este Capítulo apresenta um estudo de Revisão Sistemática da Literatura (RSL) (XIAO; WATSON, 2019) realizado com base no método proposto por Petersen et al. (2008). A busca nas bases de dados visou selecionar trabalhos publicados nos últimos cinco anos relacionados à classificação, segmentação e detecção de lesões de fundo associadas à Retinopatia Diabética e por meio da utilização de aprendizado profundo. Os artigos foram pesquisados em sete bases que são apresentadas na Tabela 2.

Tabela 2 – Bases de dados utilizadas para a realização da Revisão Sistemática da Literatura.

Biblioteca Digital	**Link**
ACM Digital Library	`https://dl.acm.org/`
IEEE Xplore	`https://ieeexplore.ieee.org`
PubMed	`https://pubmed.ncbi.nlm.nih.gov/`
Science Direct	`https://www.sciencedirect.com/`
Scopus	`https://www.scopus.com/`
Springer Link	`https://link.springer.com/`
Web of Science	`https://apps.webofknowledge.com/`

Com o objetivo de identificar o estado da arte na classificação, segmentação e detecção de lesões de fundo utilizando aprendizado profundo, a pesquisa nas bases de dados foi realizada com base nas seguintes questões de pesquisa:

- Em quais problemas relacionados ao diagnóstico de Retinopatia Diabética o aprendizado profundo é utilizado?
- Quais as técnicas/métodos/modelos/abordagens são propostas nos estudos apresentados?
- Quais são os conjuntos de dados de Retinopatia Diabética utilizados?
- Quais são os principais resultados obtidos?
- Quais são as lacunas ou questões em aberto ou trabalhos futuros relatados nos estudos?

A partir da definição dessas questões de pesquisa foram definidas as palavras-chave que foram utilizadas nas *strings* de busca. Em seguida, o mapeamento sistemático foi executado conforme as etapas descritas a seguir:

1. Busca de trabalhos correlatos utilizando *strings* de busca com base nas palavras-chave definidas, como por exemplo: (`pattern recognition` OR `deep learning` OR `convolutional neural network` OR `deep neural network`) AND (`diabetic retinopathy` OR `fundus image` OR `fundus lesions` OR `lesion detection` OR `instance segmentation`).

2. Seleção dos artigos resultantes da busca levando em consideração critérios de inclusão e exclusão. Exemplos de critérios de inclusão utilizados: “os trabalhos devem estar dentro do contexto da Retinopatia Diabética e utilizando aprendizado profundo” e “os trabalhos não devem ser mapeamentos ou revisões de literatura”. Exemplos de critérios de exclusão utilizados: “trabalhos abaixo de cinco páginas”, “trabalhos que não foram publicados em inglês ou português”, “trabalhos que não eram artigos publicados em eventos ou periódicos”, “artigos duplicados” e “trabalhos anteriores a 2017”.

3. Revisão detalhada de todos os artigos selecionados visando extrair as informações necessárias para responder as questões de pesquisa.

Após a execução do mapeamento sistemático foi realizada uma triagem dos trabalhos. Foram selecionados 51 artigos ao final das etapas 1 (busca por trabalhos correlatos) e 2 (seleção de trabalhos com base em critérios de inclusão e exclusão). Já na etapa 3, após a revisão detalhada de todos os artigos separados nas etapas 1 e 2, foram selecionados oito trabalhos.

5.1 Trabalhos Selecionados

A última etapa da Revisão Sistemática da Literatura foi a organização das características dos trabalhos relacionados que foram selecionados, apresentando o problema de pesquisa abordado no artigo, os métodos e técnicas propostos para a resolução do problema, a abordagem para avaliação do método e a indicação de limitações e trabalhos futuros, conforme é apresentado a seguir.

Benzamin; Chakraborty (2018) apresentaram uma abordagem baseada em aprendizado profundo para a detecção de exsudatos duros presentes em imagens de fundo de olho afetadas por Retinopatia Diabética. A rede neural utilizada no trabalho foi treinada com duzentos mil (200.000) *patches* de tamanho 32×32 das imagens. Os autores apresentaram uma rede neural convolucional composta por 8 camadas. A rede

prevê se o *patch* pertence à classe de exsudatos duros ou classe de fundo. Duzentas mil (200.000) imagens de treinamento foram divididas em 5 conjuntos de 40.000 imagens e cada conjunto foi treinado para 500 épocas.

A rede neural proposta pelos autores foi treinada com 1.500 épocas. Foi utilizado o otimizador Adam, e taxa de aprendizado de 0,0001. A função de perda utilizada foi Entropia Cruzada. O processo de treinamento foi realizado em mini-lotes, com 50 imagens de treinamento. A abordagem proposta por Benzamin; Chakraborty (2018) detectou exsudatos duros com uma acurácia de 98,6%. A principal limitação do trabalho foi não detectar exsudatos algodonosos, hemorragias e microaneurismas.

Já no trabalho proposto por Li et al. (2019) foi apresentado um novo conjunto de dados de Retinopatia Diabética intitulado *Dataset for Diabetic Retinopathy* (DDR) e avaliou modelos de aprendizado profundo de última geração para classificação, segmentação e detecção de lesões associadas à Retinopatia Diabética. A fim de avaliar estes métodos em situações clínicas, foram coletadas 13.673 imagens de fundo de olho de 9.598 pacientes. Estas imagens foram divididas em seis classes e avaliadas por sete especialistas para identificar o estágio de Retinopatia Diabética. Além disso, 757 imagens com Retinopatia Diabética foram selecionadas para realizar a anotação das lesões de fundo: Microaneurismas, Hemorragias, Exsudatos Duros e Exsudatos Algodonosos.

Os autores utilizaram o conjunto de dados DDR para avaliar dez modelos de aprendizado profundo de última geração, incluindo cinco modelos de classificação, dois modelos de segmentação e três modelos de detecção. Os resultados experimentais demonstram que o desempenho do modelo para o reconhecimento de microlesões deve ser melhorado para aplicar modelos de aprendizagem profunda à prática clínica. Para realizar a segmentação das lesões de RD os autores utilizaram os modelos HED (XIE; TU, 2015) e DeepLab-v3$+$ (CHEN et al., 2018), enquanto que para a detecção foram utilizados os modelos Faster R-CNN (REN et al., 2017), SSD (KONISHI et al., 2016) e YOLO (REDMON et al., 2016).

Os modelos foram avaliados com as métricas *mean Average Precision* e a média da Interseção sobre União (em inglês, *mean Intersection over Union* – $mIoU$) (MELO et al., 2020), obtidos no conjunto de validação e teste do conjunto de dados DDR. Embora os autores tenham obtido uma precisão de 0,8284 de Acurácia Geral com o modelo SE-BN-Inception para a classificação de Retinopatia Diabética, os modelos de detecção e segmentação obtiveram baixo desempenho na detecção das lesões de fundo, conforme apresentado na Tabela 3. Como trabalhos futuros, os autores visam projetar uma estrutura que combina classificação de RD com segmentação das lesões, uma vez que uma segmentação mais precisa da lesão pode melhorar os resultados obtidos com a classificação da RD.

Mateen et al. (2020) propuseram um *framework* pré-treinado baseado em uma rede neural convolucional para detectar exsudatos em imagens de fundo do olho por meio de transferência de aprendizagem. O *framework* proposto baseou-se em arquiteturas de rede neurais convolucionais pré-treinadas para detecção e classificação de exsudatos em imagens de fundo do olho.

Na estrutura proposta, três arquiteturas de redes pré-treinadas são combinadas para realizar a fusão de características, uma vez que segundo os autores, diferentes arquiteturas poderiam capturar diferentes características. Os três modelos utilizados para compor o *framework* proposto pelos autores são: Inception-v3 (SZEGEDY et al., 2016), VGG-19 (SIMONYAN; ZISSERMAN, 2015) e ResNet-50 (HE et al., 2016). Além disso, as características coletadas são tratadas como entrada nas camadas totalmente conectadas para realizar ações posteriores, como a classificação, realizada por meio da função *Softmax*. Como trabalho futuro os autores afirmam pretender estender o *framework* proposto para diagnosticar hemorragias e microaneurismas.

Já o trabalho de Wang et al. (2020) apresentou uma rede neural convolucional profunda para detecção de exsudatos duros. Os autores propuseram uma metodologia para detecção de EX usando uma rede neural convolucional profunda em conjunto com uma abordagem que utiliza morfologia matemática para segmentar EX candidatos. Após segmentar os EX candidatos e extrair automaticamente as características por meio da rede neural profunda é utilizada uma floresta aleatória para identificar os EX verdadeiros dentre todos os candidatos.

O trabalho proposto limitou-se a realizar somente a detecção de exsudatos duros. Como trabalho futuro os autores sugerem a adoção de redes neurais profundas em cascata, com estratégia de votação para melhorar o desempenho da classificação. Além disso, os autores pretendem realizar a detecção de outros tipos de lesões, como hemorragias e neovascularizações.

No trabalho de Porwal et al. (2020) são apresentados resultados de modelos de aprendizado profundo utilizados para segmentação, classificação e detecção de lesões de fundo durante o *IDRiD: Diabetic Retinopathy -– Segmentation and Grading Challenge*. A principal contribuição do trabalho foi a disponibilização do conjunto de imagens públicas de Retinopatia Diabética designado IDRiD.

Para o desafio de segmentação das lesões (microaneurismas, hemorragias, exsudatos duros e exsudatos algodonosos) a maioria das equipes participantes do desafio explorou a arquitetura U-Net (RONNEBERGER; FISCHER; BROX, 2015). Esta arquitetura é uma versão estendida das redes totalmente convolucionais (em inglês, *Fully Convolutional Network* – FCN), com o propósito de proporcionar segmentações mais precisas mesmo que em conjuntos de treino pequenos (SHELHAMER; LONG; DARRELL, 2017). Já o desafio de detecção teve por objetivo obter a localização do Disco Óptico e da Fóvea. A equipe vencedora apresentou um método baseado no modelo

Mask R-CNN para localizar e segmentar o Disco Óptico e a Fóvea simultaneamente.

Foi utilizada uma arquitetura ResNet-50 para extrair as características das imagens juntamente com uma FPN, utilizada para gerar os mapas de características em diferentes escalas, e extrair regiões de interesse das imagens. Em seguida, uma RPN percorre os mapas de características e localiza as regiões que contêm objetos. Finalmente, ramificações da arquitetura são empregadas para obter o rótulo, a máscara e a caixa delimitadora para cada região de interesse. Foi aplicada a técnica de transferência de aprendizado para treinar o modelo. Os autores iniciaram com uma taxa de aprendizado de 0,001 e um *momentum* de 0,9. A rede foi treinada com 20 épocas. Na tarefa de segmentação e detecção do Disco Óptico e Fóvea a equipe que obteve o melhor resultado no conjunto de dados de teste alcançou um coeficiente de similaridade de Jaccard igual a 0,9338.

O trabalho de Alyoubi; Abulkhair; Shalash (2021) apresentou um sistema de diagnóstico de RD. Este sistema realiza a localização das lesões na superfície da retina, sendo composto por dois modelos baseados em aprendizagem profunda. O primeiro modelo é uma CNN512, em que as imagens são utilizadas para classificar as imagens em um dos cinco estágios de RD. Foram utilizados os conjuntos de dados públicos DDR e Kaggle APTOS 2019. O segundo modelo utilizado foi uma YOLOv3, adotado para detectar as lesões RD. Finalmente, ambas as estruturas propostas, CNN512 e YOLOv3, foram combinadas para classificar imagens com RD e localizar as lesões. Não foi realizado balanceamento da quantidade de exemplos das classes de RD nem das lesões de fundo para realizar o treinamento dos modelos que compõem o sistema proposto.

A classificação do tipo de RD realizada pelo modelo CNN512 alcançou uma precisão de 88,6% e 84,1% nos conjuntos de dados públicos DDR e APTOS Kaggle 2019, respectivamente, enquanto que o modelo YOLOv3, adotado para detectar as lesões, obteve um mAP de 0,216 na detecção das lesões no conjunto de dados DDR. Como trabalhos futuros, os autores sugerem a combinação de conjuntos de dados para realizar o balanceamento da quantidade de exemplos utilizados para treinar os modelos de redes neurais profundas, e também a realização de experimentos utilizando os modelos YOLOv4 e YOLOv5.

Já o trabalho de Dai et al. (2021) apresentou um sistema de aprendizado profundo denominado DeepDR para detectar a Retinopatia Diabética usando 466.247 imagens de fundo do olho de 121.342 pacientes com Diabetes. A arquitetura do sistema DeepDR contou com três sub-redes: (1) sub-rede de avaliação da qualidade das imagens; (2) sub-rede para reconhecimento de lesões; e, (3) sub-rede para classificação de RD. Essas sub-redes foram desenvolvidas com base nos modelos ResNet e Mask R-CNN. Foram utilizadas 466.247 imagens para treinar a sub-rede de avaliação de qualidade das imagens, a fim de verificar artefatos e problemas nas imagens da

retina.

O estudo apresentou algumas limitações: a primeira foi a utilização exclusiva de um conjunto de dados de RD privado para realizar o treinamento dos modelos de aprendizado profundo, sendo utilizado o conjunto de dados de RD público Kaggle eyePACS apenas para a validação da sub-rede responsável pela classificação de RD. A segunda limitação é que a sub-rede que detecta as lesões foi testada apenas no conjunto de dados privado utilizado pelos autores devido à ausência de anotações das lesões de fundo no conjunto de dados público utilizado nos experimentos. Portanto, como trabalhos futuros os autores afirmam que uma validação adicional por meio da utilização de conjuntos de dados públicos é necessária, a fim de avaliar adequadamente o desempenho obtido pelo sistema nas tarefas de classificação de RD e detecção das lesões de fundo.

Shenavarmasouleh et al. (2021) propõem uma arquitetura para detecção de lesões de fundo composta por dois módulos. Um módulo para a detecção das lesões, e outro módulo para classificar a gravidade da RD. Os autores avaliaram a solução proposta com base nas métricas de IoU e mAP. Para realizar a detecção das lesões foi utilizado o modelo Mask R-CNN e o conjunto de dados público de RD e-Ophtha. O método utilizado pelos autores consistiu de duas etapas: (1) o conjunto de dados E-ophtha foi utilizado para treinar, ajustar e avaliar o desempenho da segmentação de instância das lesões de fundo; (2) os resultados obtidos na primeira etapa foram utilizados para identificar a gravidade da RD, com a utilização do conjunto de dados Kaggle EyePACS redimensionado. Nesta etapa, os autores utilizaram uma rede neural de memória de curto e longo prazo (em inglês, *Long Short-Term Memory* - LSTM) para realizar a identificação da gravidade da doença. Na detecção das lesões, a solução proposta foi avaliada somente no conjunto de dados e-ophtha, obtendo para o limiar de IoU de 0,5 um mAP de 0,4780 na etapa de validação, e mAP de 0,4370 na etapa de teste. O trabalho proposto limitou-se a realizar somente a detecção de exsudatos e microaneurismas.

5.2 Análise dos Trabalhos Selecionados

A Tabela 3 apresenta de forma sumarizada a comparação dos trabalhos selecionados, levando em conta critérios como métodos utilizados, objetos detectados, conjuntos de dados utilizados, principais resultados obtidos e suas limitações.

Redes neurais profundas foram aplicadas para classificar diferentes tipos e estágios de Retinopatia Diabética, segmentar e localizar as lesões associadas à Retinopatia Diabética, porém as abordagens baseadas em aprendizado profundo utilizadas apresentaram limitações nos resultados apresentados, principalmente quando aplicadas na tarefa de detecção das lesões.

Embora o aprendizado profundo apresente potencial para análise de imagens médicas, ainda possui limitações, havendo uma lacuna entre a pesquisa e a aplicação clínica. Este problema está associado com a necessidade de grandes volumes de exemplos para treinamento de abordagens baseadas em redes neurais profundas, uma vez que os conjuntos de imagens públicos de Retinopatia Diabética possuem um número reduzido de exemplos para os diferentes tipos de lesões que acometem o fundo do olho.

Também há o problema de desbalanceamento entre as classes de lesões, em que há uma incidência maior de alguns tipos de lesões em relação a outras, como no caso dos Exsudatos Algodonosos, que possuem uma quantidade menor de exemplos (anotações) nos conjuntos de dados em comparação aos Exsudatos Duros, por exemplo.

A qualidade das imagens disponibilizadas nos conjuntos de dados, assim como a presença de ruídos ou a forma que as anotações das lesões foram realizadas nestas imagens também são fatores que comprometem a precisão obtida pelas abordagens utilizadas. Outro aspecto que está associado aos baixos resultados observados nos trabalhos encontrados na literatura quanto à detecção são as características morfológicas das lesões de fundo, que assumem formatos e tamanhos variados, mesmo em lesões do mesmo tipo. Lesões como microaneurismas, por exemplo, são difíceis de detectar, principalmente em estágios iniciais da Retinopatia Diabética, em função destas lesões serem muito pequenas. Além disso, as lesões podem apresentar características diferentes de acordo com a progressão da doença para estágios mais graves.

Pode-se observar na Tabela 3 que a maioria desses trabalhos se concentrou na utilização de redes neurais convolucionais, mais especificamente de modelos de estágio único e dois estágios para a detecção de lesões em imagens de fundo. Embora estes trabalhos tenham sido publicados recentemente, as soluções propostas apresentaram limitações, como a quantidade limitada de lesões detectadas, o baixo desempenho na detecção de microlesões, o desbalanceamento dos dados utilizados para treinamento das redes neurais, e a utilização de conjuntos de dados de RD privados, que impossibilita a reprodutibilidade dos resultados. Além disso, a maioria destes trabalhos apresentou a avaliação de desempenho das abordagens propostas somente em conjuntos de dados de validação, onde os modelos foram ajustados, gerando incerteza sobre a capacidade de generalização destas soluções quando aplicadas em dados não conhecidos *a priori*.

Tabela 3 – Comparação dos Trabalhos Selecionados após a Revisão Sistemática da Literatura.

Autores	Modelos	Conjuntos de Dados	Objetos Detectados	Resultados	Principais Contribuições	Limitações
Benzamin; Chakraborty (2018)	CNN	IDRiD	EX	Acurácia = 98,6%	- Detecção de EX presentes em imagens de fundo.	- MA, SE e HE não detectados.
Li et al. (2019)	HED DeepLab-v3 Faster R-CNN SSD YOLO	DDR	EX, MA, SE e HE	mAP = 0,1702 mAP = 0,3317 mAP = 0,0924 mAP = 0,0059 mAP = 0,0035	- Disponibilização do conjunto de dados DDR para classificação de RD, segmentação e detecção de lesões de fundo. - A detecção das lesões de fundo foi realizada com base nas arquiteturas de estágio único SSD e YOLO, e com na arquitetura de dois estágios Faster R-CNN.	- Baixo desempenho na detecção e segmentação de microaneurismas e exsudatos algodonosos.
Mateen et al. (2020)	CNN	e-Ophtha DIARETDB1	EX, SE	Acurácia = 98,43% Acurácia = 98,91%	- *Framework* pré-treinado baseado em uma CNN para detectar exsudatos em imagens de fundo.	- MA e HE não detectados.
Wang et al. (2020)	CNN	e-Ophtha HEI-MED	EX	F1-*score* = 89,29% F1-*score* = 93,26%	- Uma metodologia para detecção de EX usando uma CNN em conjunto com morfologia matemática para segmentar EX candidatos.	- MA, SE e HE não detectados.
Porwal et al. (2020)	Mask R-CNN	IDRiD	Disco Óptico e Fóvea	IoU = 0,9338	- Disponibilização do conjunto de dados IDRiD para classificação de RD e segmentação de lesões. - Detecção do Disco Óptico e da Fóvea.	- EX, MA, SE e HE não detectados.
Alyoubi; Abulkhair; Shalash (2021)	CNN512 YOLOv3	Kaggle APTOS 2019 DDR	EX, MA, SE e HE	mAP = 0,216	- Um sistema de diagnóstico de RD que classifica as imagens de fundo e localiza as lesões na superfície da retina. - A detecção das lesões de fundo foi realizada com base na arquitetura de estágio único YOLOv3.	- Desbalanceamento dos dados utilizados no treinamento.
Dai et al. (2021)	ResNet Mask R-CNN	Kaggle eyePACS Privado	EX, MA, SE e HE	AUC = 0,954 AUC = 0,901 AUC = 0,941 AUC = 0,967	- Um sistema baseado em aprendizado profundo para detectar RD. - A detecção das lesões de fundo foi realizada com base na arquitetura de dois estágios Mask R-CNN.	- Conjunto de dados com imagens de fundo utilizadas para treinamento dos modelos é privado. - Validação do modelo que realiza a detecção das lesões foi realizada somente com imagens do conjunto de dados privado.
Shenavarmasouleh et al. (2021)	Mask R-CNN LSTM	e-Ophtha Kaggle eyePACS redimensionado	EX, MA	mAP = 0,4780	- Um arquitetura para detecção de lesões de fundo. - A detecção das lesões de fundo foi realizada por meio do modelo Mask R-CNN.	- Validação e teste realizado somente no conjunto de dados e-Ophtha. - SE e HE não detectados.
Abordagem Proposta para Detecção de Lesões Retinianas	YOLOv5	DDR IDRiD	EX, MA, SE e HE	mAP **= 0,2630** mAP **= 0,3280**	- Uma nova abordagem baseada em aprendizado profundo para a detecção de EX, MA, SE e HE. - Um *pipeline* para pré-processamento das imagens de fundo e aumento de dados. - Uma arquitetura de rede neural de estágio único pré-treinada e implementada com base em um modelo YOLO de última geração. - Aplicação de método para balanceamento dos dados durante o treinamento da rede neural. - Treinamento, ajuste e teste da abordagem proposta em diferentes conjuntos de dados públicos de RD.	—
Abordagem Proposta para Segmentação de Instância de Lesões Retinianas	Mask R-CNN	DDR IDRiD	EX, MA, SE e HE	mAP **= 0,2903** mAP **= 0,3063**	- Uma nova abordagem baseada em aprendizado profundo para a detecção de EX, MA, SE e HE. - Um *pipeline* para pré-processamento das imagens de fundo e aumento de dados. - Uma arquitetura de rede neural de dois estágios pré-treinada e implementada com base em um modelo de segmentação de instância de última geração. - Treinamento, ajuste e teste da abordagem proposta em diferentes conjuntos de dados públicos de RD.	—

Portanto, como a identificação de lesões retinianas em imagens de fundo possui lacunas, há um vasto espaço para novas pesquisas e propostas de novas soluções para melhorar a precisão da detecção de lesões retinianas associadas à retinopatia diabética. As abordagens para detecção e segmentação de instância de lesões de fundo desenvolvidas nesta Tese visam preencher esta lacuna propondo novas soluções baseadas em técnicas e métodos de processamento de imagens e aprendizado profundo para criação de um *pipeline* para pré-processamento das imagens de fundo e aumento de dados, e a implementação de arquiteturas de redes neurais profundas com base em modelos de última geração, cujo treino, ajuste e teste sejam realizados em diferentes conjuntos de dados públicos de RD.

5.3 Considerações sobre o Capítulo

O estudo dos trabalhos relacionados que foram identificados com a realização da revisão da literatura possibilitou identificar os métodos e as técnicas atualmente utilizadas para classificação, detecção e segmentação de objetos em imagens de fundo, bem como as métricas para avaliação destes métodos e os conjuntos de dados com imagens públicas de Retinopatia Diabética utilizados.

Os trabalhos selecionados foram apresentados com base em critérios como métodos utilizados, objetos detectados, conjuntos de dados utilizados, principais resultados obtidos e as limitações.

A partir da revisão do estado da arte foi possível verificar que as abordagens que utilizam redes neurais profundas são as mais utilizadas para identificar objetos em imagens de fundo, embora estas abordagens apresentem limitações nos resultados apresentados, principalmente quando aplicadas na tarefa de detecção das lesões de fundo.

No próximo Capítulo, é apresentada a abordagem proposta para a detecção de lesões de fundo, fundamentado tanto na apropriação conceitual realizada como na discussão relacionada ao estado da arte da área.

6 ABORDAGEM PARA DETECÇÃO DE LESÕES DE FUNDO

Este Capítulo tem por objetivo apresentar a abordagem proposta para a detecção das lesões de fundo. Serão apresentados os materiais, as técnicas e os métodos utilizados, assim como o *pipeline* usado para a construção da abordagem para a detecção das lesões de fundo. Na sequência, são apresentados e discutidos os resultados obtidos por esta abordagem durante os experimentos realizados.

6.1 Materiais, Técnicas e Métodos

Para a realização dos experimentos foi utilizado um equipamento com um Core i7 com 32 GB de RAM e uma GPU NVIDIA Titan Xp com 12 GB de VRAM. Para o processamento das imagens no ambiente de *software* foi utilizado o MATLAB[1] e o OpenCV[2] (MATUSKA; HUDEC; BENCO, 2012; ELSAYED; YOUSEF, 2019). Para a construção das estruturas responsáveis pela detecção das lesões de fundo foram utilizados o *Framework* Darknet (REDMON, 2013–2016; PUTTEMANS; CALLEMEIN; GOEDEMÉ, 2018), a *toolbox* MMdetection (CHEN et al., 2019) e a biblioteca PyTorch (PASZKE et al., 2019).

O *pipeline* da abordagem proposta está apresentado na forma de um diagrama de blocos na Figura 17. A abordagem proposta foi baseada no modelo de rede neural profunda YOLOv5 (JOCHER, 2020; ZHU et al., 2021; XU et al., 2021; QI et al., 2021; ZHU et al., 2021; RAHMAN; AZAD; HASAN, 2021; ZHENG; ZHAO; LI, 2021; XIE; ZHENG, 2021) e implementada por meio da biblioteca de aprendizado de máquina de código aberto PyTorch. Foi realizado o treinamento do modelo com 8.000 épocas e 32 lotes por época, com uma taxa de aprendizado de $0,01$ e uma taxa de *momentum* de 0,937. O tamanho das âncoras das caixas delimitadoras foi calculado de forma adaptativa (SOLAWETZ, 2020), por meio de um algoritmo genético, que otimiza as âncoras após uma varredura realizada pelo algoritmo não supervisionado K-means (COUTURIER et al., 2021) antes da etapa de treinamento.

[1] `https://www.mathworks.com/`
[2] `https://opencv.org/`

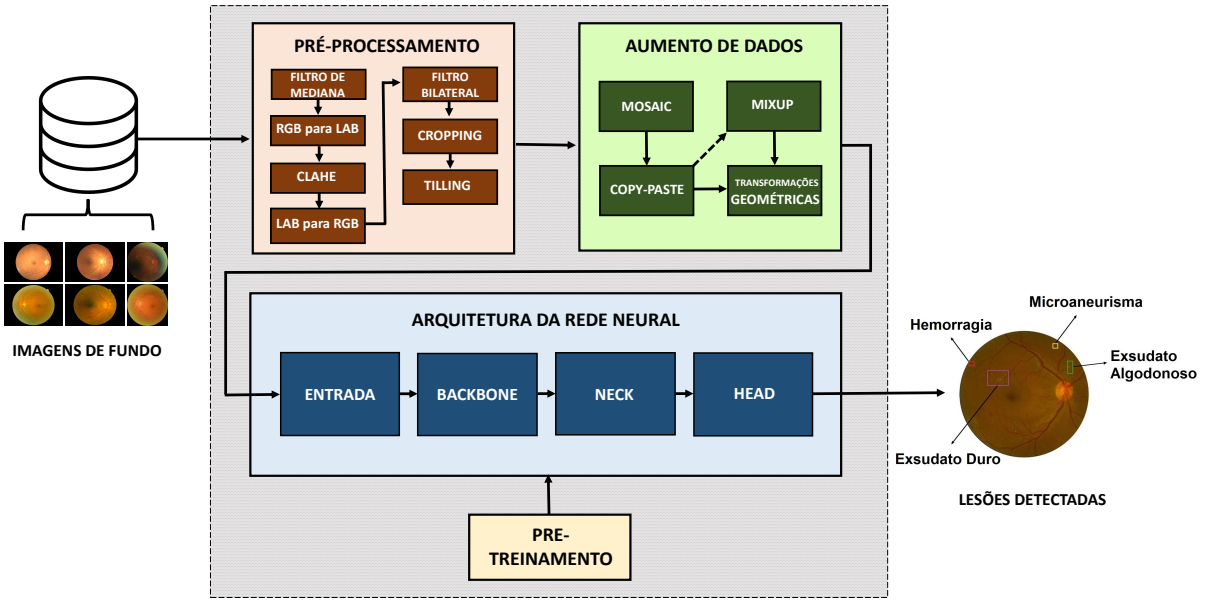


Figura 17 – Diagrama de blocos da abordagem proposta para detecção de lesões de fundo. Primeiramente, as imagens são repassadas para o bloco de Pré-processamento, para filtragem de ruídos, melhoria de contraste, eliminação parcial do plano de fundo preto das imagens e criação de *tiles*. Em seguida, as imagens pré-processadas são repassadas para o bloco de Aumento de Dados, em que são criadas artificialmente sub-imagens que serão utilizadas na camada de entrada da rede neural para treinamento da abordagem proposta, que será realizado após uma etapa de pré-treinamento da rede com os pesos ajustados no conjunto de dados Common Objects in Context.

Um conjunto mais ajustado de âncoras melhora a precisão e a velocidade das detecções (LI et al., 2021). As âncoras são os tamanhos iniciais (largura, altura) de caixas delimitadoras que serão ajustadas para o tamanho mais próximo do objeto que se pretende detectar utilizando-se alguma saída da rede neural (mapas de características final). Dessa forma, a rede não irá prever o tamanho final do objeto, mas apenas ajustar o tamanho da âncora mais próxima ao tamanho do objeto. Por esta razão, o YOLO é considerado como um método que trata a detecção de objetos como um problema de regressão, em que uma única rede neural prevê as caixas delimitadoras e as probabilidades de ocorrência de classe diretamente na imagem completa que está sendo avaliada. Como toda a detecção é realizada em uma rede (*Single-Stage*), o modelo de rede neural pode ser otimizado diretamente de ponta a ponta (REDMON et al., 2016).

Na abordagem proposta, a detecção é realizada nas camadas finais (saídas) e em três escalas, assim como proposto no modelo YOLOv3 (REDMON; FARHADI, 2018). Cada uma destas saídas ou “cabeças” de detecção possuem um conjunto separado de escalas de âncoras. No YOLOv3 são utilizados 9 tamanhos de âncoras no total, sendo 3 âncoras por cabeça de detecção. Após a detecção, alcança-se um percentual de confiança para cada lesão identificada.

YOLOv5 é um modelo de detecção de estágio único capaz de detectar objetos sem uma etapa preliminar, como no caso de detectores de dois estágios, que usam um estágio preliminar onde regiões de importância são detectadas e então classificadas para verificar se foram detectados objetos nessas áreas. A vantagem de um detector de estágio único é a velocidade com que pode fazer inferências em tempo real. Além disso, outra característica deste tipo de modelo é a possibilidade de trabalhar em dispositivos de borda e com *hardware* de baixo custo, cujo treinamento pode ser realizado com apenas uma GPU (BOCHKOVSKIY; WANG; LIAO, 2020). Pretende-se apresentar uma abordagem que alcance precisão superior às abordagens de mesmo propósito apresentados na literatura. A seguir, detalha-se metodologicamente cada etapa que compõe o *pipeline* da abordagem proposta.

6.1.1 Conjunto de Dados

Neste trabalho, usou-se o conjunto de dados de imagens DDR, que possui 757 imagens rotuladas no formato JPEG com tamanhos variáveis. Segundo (LI et al., 2019) para a captura destas imagens foram utilizadas câmeras 45 TRC NW48, Nikon D5200 e Canon CR 2, além disso, as lesões contidas nestas imagens de fundo contêm anotações (*Ground Truth*), conforme ilustrado na Figura 18.

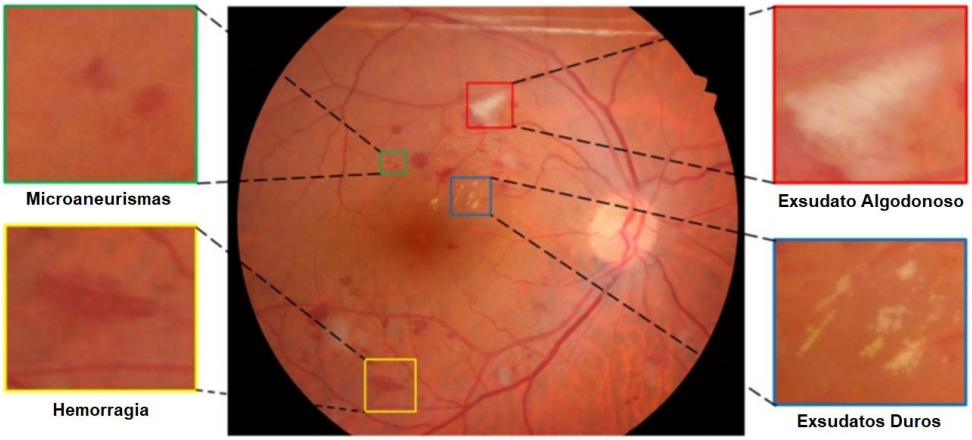


Figura 18 – Representação de uma imagem de fundo de olho com as anotações da lesão: Microaneurismas, Hemorragias, Exsudatos Algodonosos e Exsudatos Duros.

A Tabela 4 apresenta atributos do conjunto de dados DDR, como o número de imagens, a resolução das imagens, o tipo de anotações, a quantidade de imagens com anotações para MA, HE, EX e SE, e a quantidade total de anotações por tipo de lesão antes da etapa de aumento de dados.

Tabela 4 – Quantidade de imagens com anotações para MA, HE, EX e SE e a quantidade total de anotações por tipo de lesão no conjunto de dados DDR antes da etapa de aumento de dados.

# de Imagens	**Resolução**	**MA**	**HE**	**EX**	**SE**	**Anotações das lesões em nível de *pixel***	**Múltiplos especialistas**
12.522	Variável	**# de imagens com anotações por lesão**				Sim	Sim
		570	601	486	239		
		# de anotações por lesão					
		10.388	13.093	23.713	1.558		

Os dados foram coletados usando imagens de visão única. As anotações das caixas delimitadoras foram geradas automaticamente a partir das anotações em nível de *pixel* das lesões (LI et al., 2019). Embora o conjunto de dados DDR tenha boa qualidade, o treinamento da rede neural profunda da abordagem proposta possui desafios, tais como o pequeno número de lesões de fundo anotadas e a variabilidade de tamanho e forma destas lesões (vide Figura 27). Outro fator que gera problemas no treinamento de modelos de aprendizado profundo para detecção das lesões da retina é o tamanho reduzido de alguns tipos de lesões, como no caso dos microaneurismas.

Na Tabela 5, apresenta-se a definição de objetos pequenos, médios e grandes do COCO (HOSANG et al., 2016; REN et al., 2017; REZATOFIGHI et al., 2019; ZIMMERMANN; SIEMS, 2019; WANG et al., 2020; BOCHKOVSKIY; WANG; LIAO, 2020), segundo Kisantal et al. (2019). Para contornar problemas relacionados à subamostragem do conjunto de dados em função da pequena quantidade de exemplos, realizou-se um aumento de dados a partir da criação de imagens artificiais derivadas das imagens originais e anotações das lesões de fundo. Este aumento de dados e as demais técnicas que foram aplicadas para contornar os desafios supracitados serão discutidos nas Seções seguintes.

Tabela 5 – Definições de objetos pequenos, médios e grandes de acordo com o conjunto de dados COCO

Tamanho do Objeto	**Área mínima da caixa delimitadora**	**Área máxima da caixa delimitadora**
Pequeno	0×0	32×32
Médio	32×32	96×96
Grande	96×96	$\infty \times \infty$

6.1.2 Pré-Processamento das Imagens

A utilização de um bloco de pré-processamento tem como objetivos (1) melhorar a qualidade das imagens através da eliminação de padrões periódicos ou aleatórios por meio da aplicação de filtros e (2) ampliar o realce das imagens com o intuito de melhorar e acentuar as características das lesões que serão utilizadas para treinamento da

rede neural profunda. O tratamento das imagens visa (1) filtrar ruídos gerados durante a captura das imagens de fundo do olho, (2) corrigir deformidades de iluminação e (3) melhorar o contraste e realce das imagens (WALTER et al., 2002).

Para suavização da imagem, utilizou-se o filtro de mediana de tamanho 5×5. O filtro de mediana é um filtro não linear, muito eficaz na remoção de ruídos impulsivos (pulsos irregulares de grandes amplitudes), como o ruído *Salt & Pepper* (JASIM et al., 2019; GONZALEZ, R.; WOODS, 2010). Seu princípio de funcionamento consiste em substituir o valor de um *pixel* pela mediana dos níveis de intensidade na vizinhança desse *pixel* (GONZALEZ, R.; WOODS, 2010). Para realizar a filtragem de mediana é necessária a especificação do tamanho do filtro para que sejam verificados os valores de *pixel* abrangidos por este filtro, a fim de determinar o nível mediano destes *pixels* (TAN; JIANG, 2019).

O filtro de mediana é utilizado em processamento de imagens para remoção de ruído com a preservação das bordas (JASIM et al., 2019). A vantagem deste filtro está em manter a resolução espacial e os principais detalhes da imagem enquanto pontos isolados são removidos (FARIA, 2010). Este filtro proporciona a redução de ruído com menos borramentos em comparação com filtros lineares de suavização similares (GONZALEZ, R.; WOODS, 2010). No hiperespaço de uma imagem colorida, conforme apresentado na Equação 5, classificar os valores de intensidade de um conjunto de *pixels* é equivalente a classificar as normas dos vetores de intensidade, conforme Equação 6 (ELHEDDA; MEHRI, 2019).

$$I(x-1:x+1, y-1:y+1) = \begin{pmatrix} a & b & c \\ d & e & f \\ g & h & i \end{pmatrix} \tag{5}$$

onde $a, b, \ldots, h, i$ representam um conjunto de vetores de modo que cada vetor contém diferentes componentes de cores do espaço de cores selecionado.

$$F(x,y) = median(||a||, ||b||, \ldots, ||h||, ||i||) \tag{6}$$

Na Figura 19 apresentam-se os resultados obtidos com a aplicação do Filtro de Mediana com filtro de tamanho 5×5 em uma imagem de fundo do olho do conjunto de dados DDR após a introdução artificial dos ruídos *Salt & Pepper, Gaussian* e *Speackle* (JASIM et al., 2019), para fins de validação da eficácia do filtro.

Analisaram-se as imagens por meio da métrica $PSNR$ (*Peak Signal-To-Noise Ratio*) (FARDO et al., 2016; ERFURT et al., 2019), tal que quanto maior o $PSNR$ melhor será a qualidade da imagem compactada ou reconstruída (SHIAO et al., 2007). Portanto, o $PSNR$ pode ser usado para comparar a qualidade de compressão da imagem. Para calcular o $PSNR$, primeiramente é calculado o *Mean Squared Error* (MSE)

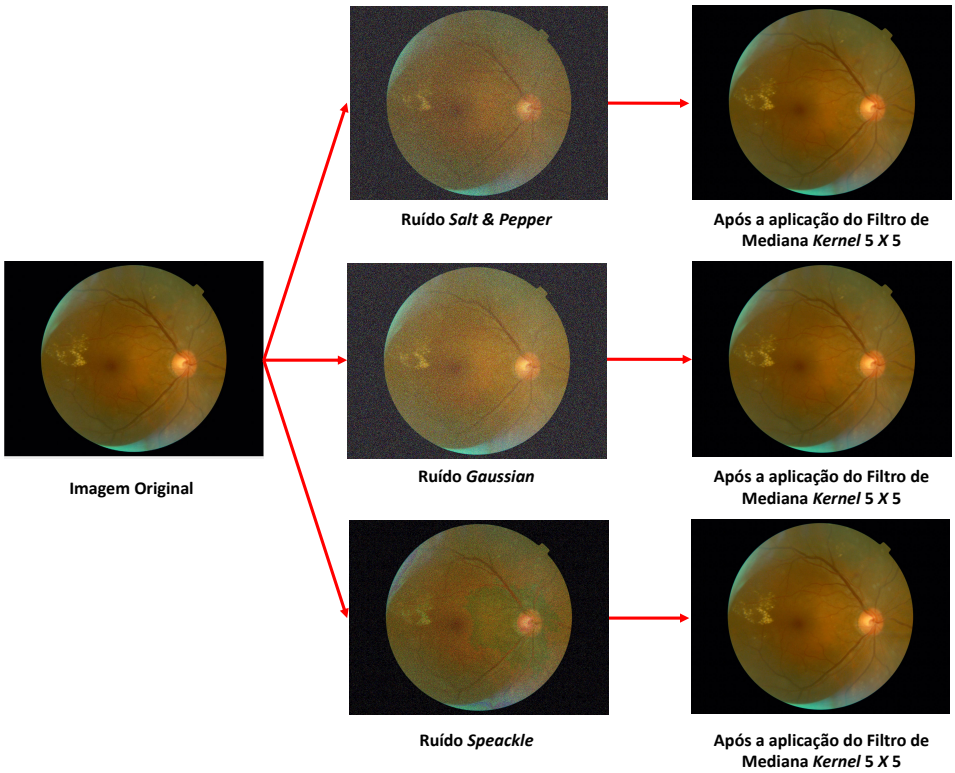


Figura 19 – Remoção dos ruídos do tipo *Salt & Pepper*, *Gaussian* e *Speackle* com a aplicação do Filtro de Mediana com *kernel* de tamanho 5×5 em uma imagem de fundo do conjunto de dados DDR.

usando a Equação 7.

$$MSE = \frac{\sum_{m,n}[I_1(m,n) - I_2(m,n)]^2}{M * N} \tag{7}$$

onde M e N são o número de linhas e colunas nas imagens de entrada. Em seguida, o $PSNR$ é calculado usando a Equação 8.

$$PSNR = 10log_{10}\left(\frac{R^2}{MSE}\right) \tag{8}$$

onde, R é a flutuação máxima no tipo de dados da imagem de entrada. Por exemplo, se a imagem de entrada tiver um tipo de dados de ponto flutuante de precisão dupla, R será 1. Se tiver um tipo de dados inteiro sem sinal de 8 bits, R será 255.

A Tabela 6 apresenta os resultados obtidos com a métrica $PSNR$ após a remoção dos ruídos do tipo *Salt & Pepper*, *Gaussian* e *Speackle* com a aplicação do Filtro de Mediana de tamanho 5×5 em uma imagem de fundo do conjunto de dados DDR. Por meio da análise objetiva obtida com o $PSNR$, verificou-se que o Filtro de Mediana obteve bons resultados na redução dos três principais tipos de ruídos, principalmente no *Gaussian* e no *Salt & Pepper*.

Tabela 6 – Resultados obtidos com a métrica $PSNR$ após a remoção dos ruídos do tipo *Salt & Pepper*, *Gaussian* e *Speackle* com a aplicação do Filtro de Mediana com *kernel* de tamanho 5×5 em uma imagem de fundo do conjunto de dados DDR.

Tipos de ruído	$PSNR$ **- Imagem com ruído**	$PSNR$ **- Filtro Mediana *kernel* 5×5**
Salt & Pepper	8,0949	12,9978
Gaussian	7,9777	12,9978
Speackle	11,4344	12,9978

Para realizar o realce das imagens foi utilizada a técnica de equalização adaptativa de histograma limitado por contraste (CLAHE) (SANTOS et al., 2021; ALYOUBI; ABULKHAIR; SHALASH, 2021). Esta técnica foi desenvolvida inicialmente para realce de imagens de baixo contraste e é uma evolução do método de equalização de histograma (RAI; GOUR; SINGH, 2012) e tem sido usada como parte de *pipelines* de pré-processamento para melhorar a qualidade de imagens médicas (HORRY et al., 2020). O algoritmo divide a imagem em pequenas regiões (blocos) e aplica a equalização do histograma em cada uma dessas regiões, conforme ilustra-se na Figura 20. A vantagem dessa técnica é considerar a equalização local da imagem, com a definição de um limite de nível de cinza.

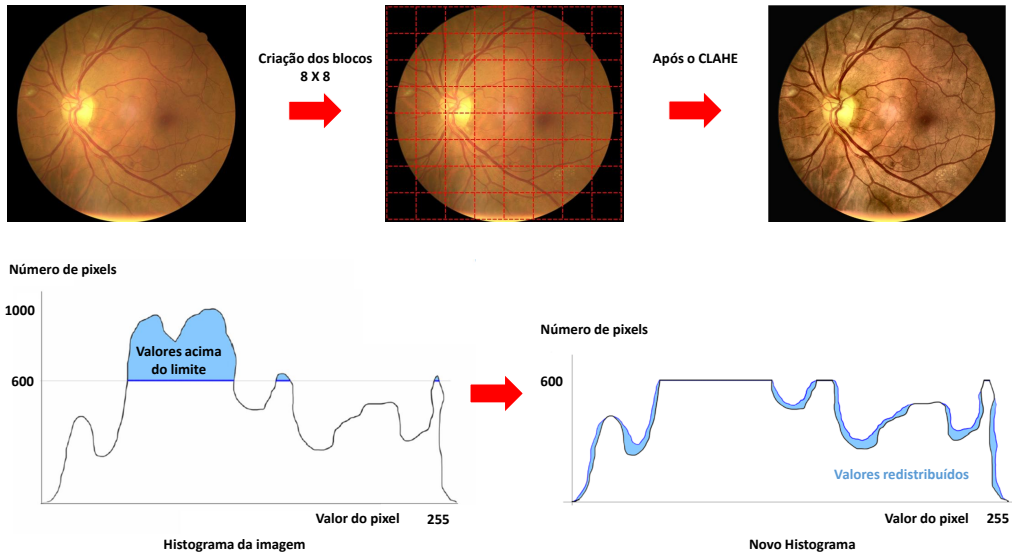


Figura 20 – CLAHE divide a imagem em blocos (*tiles*) de tamanho 8×8 e aplica a equalização do histograma em cada uma dessas regiões fazendo com que os valores dos *pixels* que estão acima de um limiar pré-definido sejam redistribuídos em um novo histograma.

Ao contrário de técnicas convencionais de equalização de histograma que opera em uma imagem como um todo, o CLAHE é um método que processa pequenas regiões da imagem e os combina com interpolação bilinear para evitar a geração de artefatos (MUKHOPADHYAY et al., 2015). Assim como nos trabalhos de Park; Cho; Choi (2008), Setiawan et al. (2013), Yadav; Maheshwari; Agarwal (2014), e D. D. Silva; B. P. Carneiro; F. S. Cardoso (2018), aplicou-se o CLAHE com blocos de tamanho 8×8 para

realizar a equalização do histograma em cada uma dessas regiões fazendo com que os valores dos *pixels* que estão acima de um limiar pré-definido fossem redistribuídos em um novo histograma. O pseudocódigo do CLAHE é descrito no Algoritmo 1 (YADAV; MAHESHWARI; AGARWAL, 2014).

Algoritmo 1 Pseudocódigo do CLAHE

1: Processar a aquisição da imagem.
2: Definir a configuração dos parâmetros ▷ Número de *tiles*; número de *bins* (compartimentos de histograma usados para construir uma transformação de realce de contraste) usados no histograma; *Clip Limit* (limite de intensificação de contraste); tipo de distribuição (forma de apresentação do histograma), etc.
3: Dividir a imagem original em *tiles*.
4: Pre-processar cada *tile* obtido na etapa anterior.
5: Gerar o mapeamento do nível de cinza e realizar o corte no histograma. ▷ No *tile*, o número de *pixels* são divididos igualmente em cada nível de cinza, então o número médio de *pixels* em nível de cinza é obtido de acordo com a Equação 9.
6: Interpolar o mapeamento de nível de cinza para criar a imagem aprimorada.

$$N_{avg} = \frac{N_{CR-Xp} * N_{CR-Yp}}{N_{gray}} \tag{9}$$

onde: N_{avg} indica o número médio de *pixels*; N_{gray} é o número de nível de cinza dos *tiles*; N_{CR-Xp} é o número de *pixels* na direção x dos *tiles*; N_{CR-Yp} é o número de *pixels* na direção y dos *tiles*.

Após, o cálculo do *Clip Limit* real é realizado de acordo com a Equação 10.

$$N_{CL} = N_{CLIP} * N_{avg} \tag{10}$$

No entanto, antes de se aplicar o algoritmo CLAHE nas imagens de fundo do conjunto de dados foi necessário definir qual seria o espaço de cores mais adequado para realizar o realce das imagens. A Figura 21 apresenta uma ilustração dos diferentes espaços de cores utilizados nos experimentos do trabalho proposto, em que à esquerda está o modelo RGB (na forma de um cubo); no centro, o modelo HSV (na forma de cone); e à direita, o modelo LAB (na forma de esfera).

As imagens do conjunto de dados DDR estão no espaço de cores RGB. Em um modelo RGB, uma imagem digital consiste em três planos de imagens, cada uma das quais armazena os valores de R (*Red*), G (*Green*) e B (*Blue*) (MA et al., 2018). RGB é ideal para gerar imagens, por monitores ou câmeras, por exemplo, porém os efeitos de aprimoramento neste espaço de cores, como contraste e informação de entropia são muito limitados (MA et al., 2018). Os três canais do RGB têm uma forte correlação porque todos contêm intensidade luminosa em sua formação, sendo assim, a aplica-

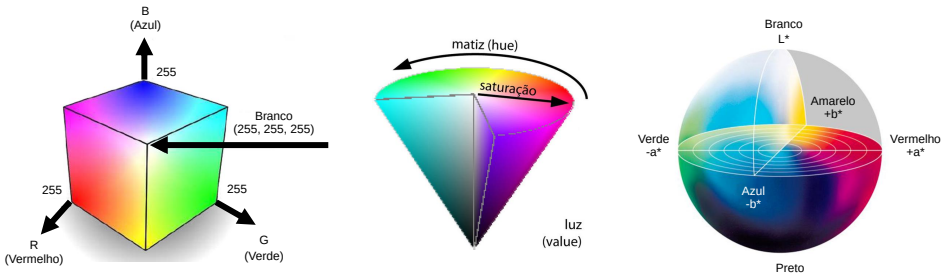


Figura 21 – Ilustração dos diferentes espaços de cores utilizados nos experimentos do trabalho proposto. À esquerda, o modelo RGB (formato cúbico); no centro, o modelo HSV (formato cônico); e, à direita, o modelo LAB (formato esférico).

ção de realce diretamente nestes componentes geralmente não consegue alcançar resultados desejados (DAI; FAN; PENG, 2018).

Há uma dificuldade em especificar uma cor através de três cores primárias, já que as informações de cores e intensidade luminosa (brilho) estão juntas, dificultando processos em que essas componentes precisam ser analisadas separadamente. Para contornar este problema, primeiramente, as imagens originais foram divididas em três imagens R, G e B independentes. Na sequência, aplicou-se o algoritmo CLAHE somente no canal G, pois lesões como microaneurismas apresentam um melhor contraste neste canal em comparação aos canais R e B (D. D. SILVA; B. P. CARNEIRO; F. S. CARDOSO, 2018).

O espaço de cores HSV é baseado na teoria da percepção visual humana e é adequado para descrever e interpretar cores (LIU et al., 2012). O modelo HSV define uma cor no espaço em termos de componentes H (*Hue*), S (*Saturation*) e V (*Value*). Este espaço de cores dissocia informações acromáticas (componente V, valor relacionado ao brilho), a partir de informações cromáticas (componentes H e S), em uma imagem colorida (GONZALEZ; WOODS; EDDINS, 2003). Comparado com o espaço de cores RGB, o espaço de cores HSV é mais próximo da percepção de cor do olho humano. Outra vantagem, é que o componente relacionado ao brilho pode ser modificado independentemente dos demais componentes, sendo menos sensível a ruídos (ZHAO; LI; FENG, 2008). Sua importância se dá pela possibilidade de analisar e/ou modificar os níveis de intensidade do brilho presente na imagem independente dos canais H e S. Para se aplicar o algoritmo CLAHE neste espaço de cores, primeiramente, converteu-se a imagem original para o espaço HSV e, em seguida, fez-se o realce somente no canal V, relacionado à intensidade do brilho da imagem.

O espaço de cores LAB tem ampla gama de cores, pode expressar todas as cores percebidas pelo olho humano e compensa o problema de distribuição do RGB (DAI; FAN; PENG, 2018). O espaço de cores LAB também conhecido como CIE-LAB, sendo definido pela *International Commission on Illumination* (CIE) em 1976, cuja composi-

ção tem uma dimensão (L), relacionada à luminosidade, e as dimensões A e B relacionadas às cores (WARNER, 2014).

LAB é baseado no modelo de cor oponente da visão humana, onde vermelho e verde formam um par oponente e azul e amarelo formam outro par oponente (PUJARI; PUSHPALATHA; PADMASHREE, 2010; WARNER, 2014). Além disso, esse espaço de cores é muito próximo ao sistema visual humano, fazendo com que hajam mais informações em comparação ao RGB, por exemplo (SINGH; TIWARI, 2018). Sua principal vantagem é permitir que tons e cores sejam balanceados de forma interativa e independente. O espaço de cores LAB tem as cores uniformemente distribuídas, sendo o canal Luminosidade (L) separado da cromaticidade (A e B). A utilização deste espaço de cores é a maneira mais precisa de chegar em uma cor exata, pois é possível reproduzir todas as cores existentes no espectro visível. Para se aplicar o algoritmo CLAHE neste espaço de cores, primeiramente, converteu-se a imagem original para o formato LAB, e, em seguida, fez-se o realce somente no canal L, relacionado à luminosidade da imagem.

Nos experimentos, assim como no trabalho proposto por Alyoubi; Abulkhair; Shalash (2021), utilizou-se um tamanho de *tiles* 8$\times$8 (ALYOUBI; ABULKHAIR; SHALASH, 2021) com *Clip Limit* igual a 6 em todos os espaços de cores investigados. Também usou-se *bins* igual a 300 e *distribution* igual a *rayleigh*. O realce de uma imagem tem por propósito aumentar o contraste de objetos que possuem baixo contraste objetivando uma melhor visualização destes. No entanto, mensurar qualitativamente o realce obtido em uma imagem não é algo simples (D. D. SILVA; B. P. CARNEIRO; F. S. CARDOSO, 2018), pois este tipo de avaliação varia de pessoa para pessoa, inclusive com critérios subjetivos (SANTOS, 2016). Dessa forma, existem métricas objetivas que podem auxiliar a estimar o contraste de forma quantitativa (SCHETTINI et al., 2010). No entanto, estas métricas não determinam necessariamente se a imagem possui boa ou má qualidade com relação ao realce, mas frequentemente explicam características importantes da imagem (WANG et al., 2013).

Para aferir quantitativamente o melhor contraste obtido nas imagens após o realce utilizaram-se as métricas *Measure of Enhacement* (EME) (LENTZ; GRIGORYAN, 2000) e Entropia (E) (YE; MOHAMADIAN; YE, 2007). Inicialmente, calculou-se a Entropia das imagens de fundo a fim de avaliar a melhoria de realce após a aplicação do algoritmo CLAHE. A entropia é uma medida estatística de aleatoriedade que mede a informação média de um resultado aleatório. No caso de avaliação de imagens, altos valores significam que todos os níveis de cinza possuem a mesma probabilidade (SANTOS, 2016). Portanto, esta medida traz uma indicação sobre o nível de contraste da imagem, já que esta informação está relacionada com a forma de distribuição dos *pixels* ao longo do histograma. Desse modo, sendo $P(x_i)$ a probabilidade do nível de cinza i, define-se a entropia $E(x)$ pela Equação 11 (YE; MOHAMADIAN;

YE, 2007).

$$E(x) = -\sum_{i=0}^{255} P(x_i) \log P(x_i) \tag{11}$$

Portanto, esta métrica fornece um indicativo do contraste global da imagem, ao qual valores altos retratam o alargamento do histograma e o aumento da uniformidade da frequência de cada *pixel* (SANTOS, 2016). Após a verificação da Entropia das imagens calculou-se o EME das imagens realçadas para verificar quais os melhores resultados obtidos nos espaços de cores RGB, HSV e LAB. EME foi uma métrica proposta originalmente no trabalho de Lentz; Grigoryan (2000), para calcular a qualidade do realce em imagens digitais.

O método utilizado pela métrica EME consiste em dividir a imagem em uma matriz de sub-imagens w, onde cada sub-imagem é uma matriz quadrada. O ideal é que esta sub-matriz tenha a mesma dimensão dos *tiles* (sub-matrizes nas quais o método CLAHE divide a imagem para aplicar o realce). Após dividir a imagem realçada em sub-matrizes, para cada sub-matriz é calculada a razão entre os *pixels* que possuem maior e menor intensidade de cinza. Em seguida, calcula-se a magnitude da razão, gerando assim, um valor EME para cada sub-matriz. Por fim, o valor EME da imagem é obtido calculando a média dos valores EME obtidos em cada sub-matriz, quantificando desta forma, a melhoria de contraste da imagem após o realce, conforme a Equação 12 (D. D. SILVA; B. P. CARNEIRO; F. S. CARDOSO, 2018).

$$EME = \frac{1}{K_1 K_2} \sum_{l=1}^{K_2} \sum_{k=1}^{K_1} 20 \log \left(\frac{I_{MAX}^W; k, l}{I_{MIN}^W; k, l} \right) \tag{12}$$

onde: K_1 indica a quantidade de linhas na matriz de sub-imagens; K_2 a quantidade de colunas na matriz de sub-imagens; I_{MAX}^W a intensidade máxima de nível de cinza na sub-imagem w; e, I_{MIN}^W a intensidade mínima de nível de cinza na sub-imagem w.

Esta medida está relacionada ao contraste local de modo que valores altos indicam alto contraste local, enquanto valores próximos de zero indicam regiões homogêneas (SANTOS, 2016). Para aplicação da métrica EME utilizou-se como tamanho de sub-blocos (sub-imagens) o tamanho de 8$\times$8, mesmo valor apresentado nos trabalhos de Huynh-the et al. (2014), Santos (2016) e D. D. Silva; B. P. Carneiro; F. S. Cardoso (2018). Após o realce das imagens de fundo com CLAHE nos espaços de cores RGB(G), HSV(V) e LAB(L), teve-se que converter estas imagens para níveis de cinza para se aplicar a métrica EME e avaliar a melhoria de contraste obtido em cada espaço de cores. Estas imagens em níveis de cinza foram utilizadas apenas para fins de avaliação objetiva com a métrica EME.

Com o algoritmo CLAHE aplicado para realçar as imagens de fundo do olho nos espaços de cores RGB, HSV e LAB verificaram-se melhorias no contraste das lesões, tanto pela avaliação subjetiva das imagens e seus histogramas, como pela análise objetiva realizada por meio das métricas de Entropia e *Measure of Enhancement*. Percebeu-se que o aumento significativo de Limiar (*Clip Limit*) na parametrização do CLAHE gerou artefatos em algumas regiões das imagens. A avaliação objetiva do contraste obtido em 15 imagens de fundo do conjunto DDR é apresentada nas Tabelas 7 e 8.

Tabela 7 – Resultados obtidos com a métrica Entropia após a aplicação do algoritmo CLAHE nas imagens de fundo do conjunto de dados DDR nos espaços de cores RGB(G), HSV(V) e LAB(L).

Imagens	E **CLAHE + RGB(G)**	E **CLAHE + HSV(V)**	E **CLAHE + LAB(L)**
#007-1974-100	5,3406	4,8777	4,7159
#007-3319-200	4,5547	4,5681	4,1461
#007-3336-200	6,2273	5,8598	5,9846
#007-3338-200	6,2343	5,5323	5,9828
#007-3387-200	7,2854	6,7330	5,9985
#007-3396-200	5,4818	5,4452	4,9015
#007-3427-200	6,0415	5,7405	5,0914
#007-3939-200	6,2916	5,8231	5,7162
#007-4273-200	5,4263	5,7404	4,5563
#007-6335-400	5,9840	5,7604	4,7208
#007-6361-400	5,2315	4,8563	4,5155
#007-6686-400	6,8045	6,3330	6,2147
#007-6717-400	6,7205	6,3416	5,7634
#007-6722-400	6,6873	6,4085	6,1214
#007-6926-400	5,3639	5,3364	5,1495

Com base nestes resultados, na análise dos histogramas e das imagens geradas após a aplicação do CLAHE, é possível afirmar que os melhores resultados de realce foram obtidos nos espaços de cores RGB(G) e LAB(L), respectivamente, sendo que as imagens no espaço de cores LAB produziram os melhores resultados, mantendo um bom nível de realce, conforme pode ser observado na Figura 22.

Assim, optou-se pelo espaço de cores LAB para fazer o realce porque as imagens coloridas apresentam consigo características das lesões de fundo que poderiam ser perdidas caso optássemos pelas imagens no canal (G) do RGB. A aplicação do pré-processamento com o algoritmo CLAHE gerou imagens de melhor qualidade, ajustando o contraste e reduzindo ruídos. No entanto, mesmo com o ajuste do limiar mais adequado para as imagens de fundo, foi observada a geração de alguns artefatos após a aplicação do CLAHE. Nesse sentido, a próxima etapa de pré-processamento foi a aplicação de um filtro para suavização das imagens, com o propósito de minimizar es-

Tabela 8 – Resultados obtidos com a métrica EME após a aplicação do algoritmo CLAHE nas imagens de fundo do conjunto de dados DDR nos espaços de cores RGB(G), HSV(V) e LAB(L).

Imagens	EME **CLAHE + RGB(G)**	EME **CLAHE + HSV(V)**	EME **CLAHE + LAB(L)**
#007-1974-100	3,5948	3,2152	2,1602
#007-3319-200	2,1958	2,0634	1,1713
#007-3336-200	3,0931	2,9379	1,9840
#007-3338-200	3,5287	2,6052	2,2821
#007-3387-200	3,7507	3,6807	2,0064
#007-3396-200	3,9257	3,5753	2,5867
#007-3427-200	3,9591	4,2571	3,2447
#007-3939-200	2,9237	2,7346	1,6858
#007-4273-200	2,1647	1,9182	1,1121
#007-6335-400	3,7807	4,5371	2,5995
#007-6361-400	3,1811	3,0705	1,7046
#007-6686-400	6,1646	6,0830	4,3246
#007-6717-400	7,0138	6,9563	4,2578
#007-6722-400	6,7473	6,5823	4,0060
#007-6926-400	4,2441	3,9220	3,2276

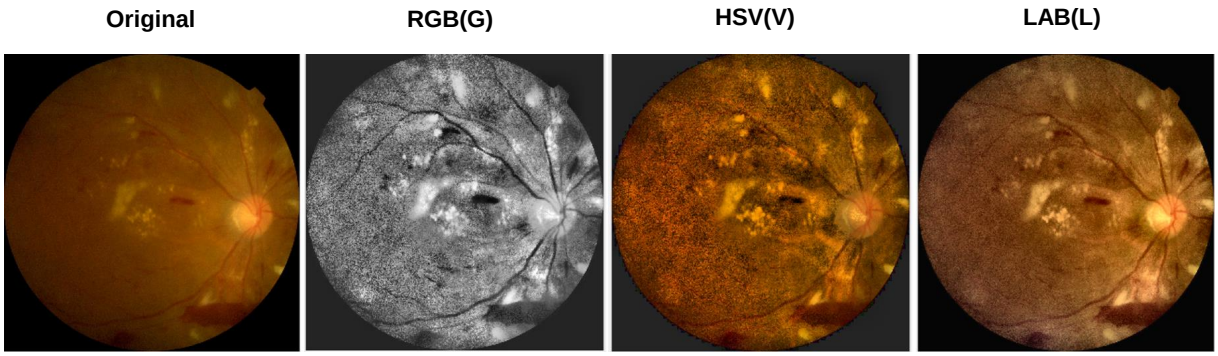


Figura 22 – Exemplo de contrastes obtidos com o algoritmo CLAHE nos canais (G) do RGB, (V) do HSV, e (L) do LAB, usando tamanho de *tile* de 8×8 e *Clip Limit* igual a 6.

tes artefatos que foram criados após o realce sem, no entanto, perder características importantes das imagens, tais como as bordas das lesões que pretendemos detectar.

Após o realce das imagens de fundo com CLAHE converteu-se as imagens para o padrão RGB e aplicou-se o Filtro Bilateral para prover a suavização das imagens. O Filtro Bilateral é um filtro de suavização não linear que preserva os detalhes de bordas (YANG; TAN; AHUJA, 2009; PARIS; DURAND, 2006; SUN et al., 2011). O filtro leva em consideração a distância entre os *pixels* no espaço (chamada similaridade de proximidade espacial) e, também, considera o grau de semelhança entre *pixels* (chamada de escala de cinza similaridade) (DAI; FAN; PENG, 2018). Este filtro é semelhante ao Filtro Gaussiano pois encontra a média ponderada gaussiana na vizinhança, porém leva em conta a diferença de *pixel* ao borrar os *pixels* próximos. O

Filtro Bilateral é representado pela Equação 13 (DAI; FAN; PENG, 2018).

$$BF(I) = \left\{ \sum_{n \in \Omega} w(x,n) g(n) \right\} * \left\{ \sum_{n \in \Omega} w(x,n) \right\}^{-1} \tag{13}$$

onde: $BF(I)$ é uma imagem usando filtragem bilateral; $w(x,n)$ é função de peso da filtragem bilateral; $g(n)$ é o *haze image*; e, Ω é a área local com o x como o centro. A função de peso pode ser expressa como:

$$w(x,n) = w_{\sigma_d} * w_{\sigma_r} \tag{14}$$

onde: w_{σ_d} é a função do kernel do espaço; w_{σ_r} é o valor do kernel da função de domínio; σ_d é a variação da influência da similaridade de proximidade espacial; e, σ_r é a variância dos fatores de influência da similaridade em escala de cinza. w_{σ_d} e w_{σ_r} podem ser expressos como:

$$w_{\sigma_d} = \exp\left[-\frac{d_d^2(x,(x-n))}{2\sigma_d^2} \right] = \exp\left[-\frac{n^2}{2\sigma_d^2} \right] \tag{15}$$

$$w_{\sigma_r} = \exp\left[-\frac{d_d^2(x,(x-n))}{2\sigma_r^2} \right] = \exp\left[-\frac{(g(x)-g(x-n))^2}{2\sigma_r^2} \right] \tag{16}$$

Embora o Filtro Bilateral tenha um tempo de processamento superior aos demais filtros de suavização, tem como vantagem garantir que apenas os *pixels* com intensidade semelhante ao *pixel* central fiquem desfocados, enquanto os demais *pixels* com valores distintos não ficam desfocados. Ao fazer isso, as arestas que possuem maior variação de intensidade são preservadas. Sendo assim, o Filtro Bilateral preserva as bordas, pois os *pixels* nas bordas terão grande variação de intensidade.

Utilizou-se um tamanho de filtro $d = 9$ (diâmetro de cada vizinhança de *pixel* durante a filtragem) para que fosse realizada uma filtragem de ruídos mais profunda e $\sigma = 75$ para que as lesões presentes nas imagens de fundo não ficassem desfocadas. Para fins de mensuração, verificou-se o $PSNR$ de uma imagem de fundo realçada com o CLAHE em que obtivemos o resultado de 18,9808. Após a suavização desta mesma imagem com o Filtro Bilateral obtivemos um $PSNR$ de 19,0772. Realizou-se a verificação do $PSNR$ das demais imagens e constatou-se que após a suavização com o Filtro Bilateral obteve-se uma melhoria na relação sinal-ruído de pico das imagens, assim como uma redução de *outliers* provenientes do processo de realce.

Na Figura 23 apresentam-se imagens de fundo de olho do conjunto de dados DDR, antes e depois da etapa de pré-processamento e preparação, onde à esquerda está a imagem original e seu histograma em nível de cinza e à direita está a imagem pré-processada com seu histograma no nível de cinza.

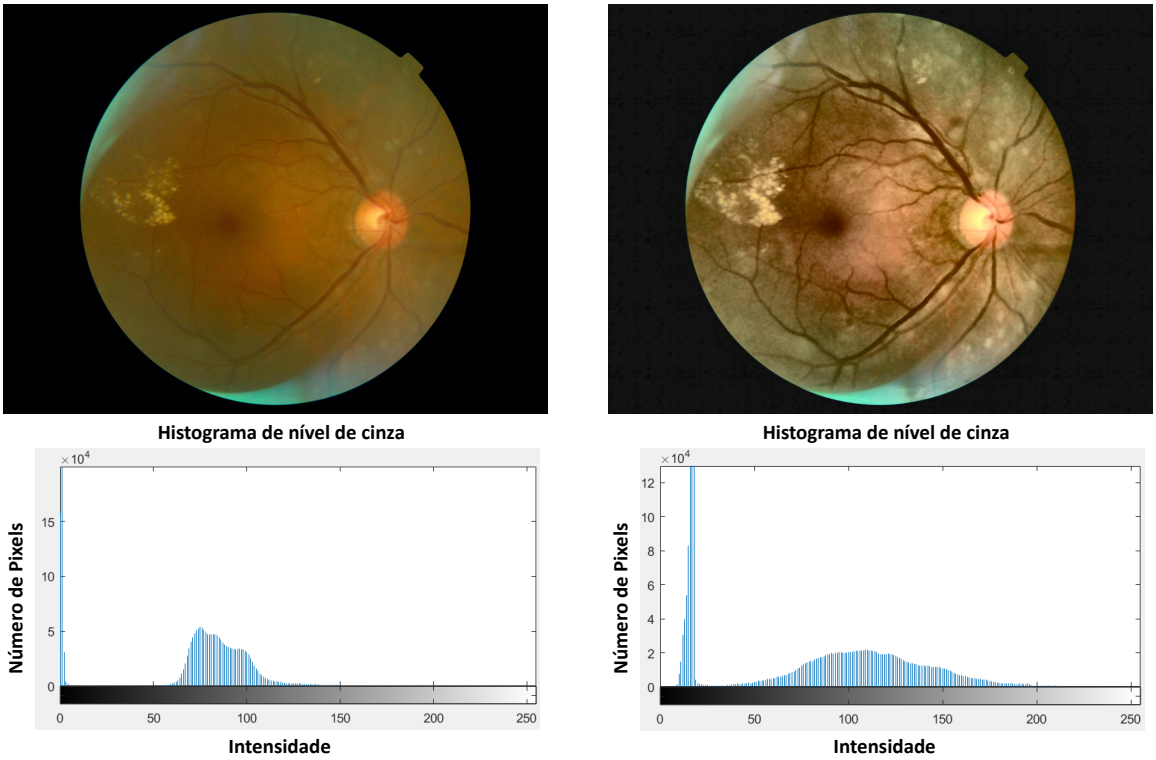


Figura 23 – Imagem de fundo do olho do conjunto de dados DDR antes e após a etapa de Pré-Processamento e Preparação. À esquerda imagem original e seu histograma em nível de cinza e à direita a imagem pré-processada com seu histograma em nível de cinza.

Assim como os trabalhos de El abbadi; Hammod (2014) e de Alyoubi; Abulkhair; Shalash (2021), realizou-se uma etapa de pré-processamento para *cropping* parcial do plano de fundo preto das imagens da retina. Segundo El abbadi; Hammod (2014), a importância de remover o plano de fundo preto da imagem da retina está relacionado à geração de falsos positivos durante a detecção das lesões, principalmente na fronteira da retina, onde há uma semelhança da borda da retina com os vasos sanguíneos. No caso das imagens de fundo do olho, apenas os *pixels* da retina possuem informações significativas, o restante é considerado plano de fundo sendo, portanto, importante realizar a localização da área de interesse e remoção das características não desejadas relacionadas ao plano de fundo preto da imagem.

Primeiramente, realizou-se a detecção da retina nas imagens de fundo por meio da Transformada de Hough (HT), uma técnica que localiza formas em imagens, sendo usada para extrair linhas, círculos e elipses (ou Seções cônicas) (NIXON; AGUADO, 2020). HT é um método amplamente utilizado para detecção e reconhecimento de curvas devido à sua robustez e capacidade de processamento (RONG; DU-WU; BO, 2009), sendo tipicamente utilizado para detectar ou segmentar objetos geométricos a partir de imagens (YE et al., 2015). HT é um método popular em Visão Computacional (CHANDRASEKAR; DURGA, 2014), para detecção de formas que são facilmente parametrizadas (linhas, círculos, elipses, etc) em imagens computacionais. Em geral, esta transformada é aplicada após a imagem passar por um pré-processamento,

comumente a detecção de bordas (YE et al., 2015).

A Transformada de Hough aplica uma transformação na imagem tal que todos os pontos pertencentes a uma mesma curva são mapeados em um único ponto de um espaço de parâmetros, também chamado de espaço de Hough (PEIXOTO, 2003). Cada borda de uma imagem é transformada pelo mapeamento para determinar células no espaço de parâmetros, indicadas pelas primitivas definidas através do ponto analisado. Essas células são incrementadas e, por meio da máxima local do acumulador no final do processo, indicará os parâmetros correspondentes à forma especificada (YE et al., 2015). A HT é aplicada aos dados de um mapa de bordas obtido na etapa de segmentação da imagem e permite detectar praticamente qualquer tipo de curva, mesmo aquelas pouco visíveis e fortemente ruidosas (CASTRO, 1996).

Embora a HT tenha sido introduzida para detecção de retas, é possível generalizá-la para detectar círculos, elipses ou ainda uma curva qualquer parametrizável na forma $h(v, c)$, onde v é o vetor de coordenadas e c é o vetor de parâmetros. A desvantagem da utilização da HT é o esforço computacional decorrente do aumento da dimensionalidade do espaço de parâmetros, pois estas curvas já não poderão mais ser representadas num espaço bidimensional (PEIXOTO, 2003). Seja λ uma circunferência de centro (a, b) e raio r do plano $x - y$ representada pela Equação 17:

$$\lambda : (x - a)^2 + (y - b)^2 = r^2 \tag{17}$$

Através da HT um ponto (x_i, y_i) pertencente a λ pode ser transformado numa superfície cônica λ' do plano paramétrico tridimensional a, b, r, definida pela Equação 18:

$$\lambda : (a - x_i)^2 + (b - y_i)^2 = r^2 \tag{18}$$

Assim como o espaço paramétrico, os acumuladores também serão tridimensionais para as configurações circulares na imagem. Assim, será construída a matriz de acumuladores, onde um valor N numa célula (a_i, b_j, r_k) indicará que N pontos pertencem à circunferência λ' no plano x, y definida por $(x - a_i)^2 + (y - b_j)^2 = (r_k)^2$. A extração de circunferências existentes na imagem é efetuada através dos acumuladores de maior peso, os quais representam os pontos com maior número de interseções entre os cones paramétricos (PEIXOTO, 2003).

Antes de identificar o contorno da retina, realizou-se o pré-processamento aplicando-se o Filtro de Mediana para suavizar as imagens e eliminar detalhes irrelevantes para a detecção da circunferência da retina. Depois, realizou-se a detecção dos contornos por meio da HT. Quando a forma da retina é encontrada na imagem de fundo, então o mapeamento de todos seus pontos no espaço de parâmetros se agrupam em torno dos valores de parâmetros que correspondem à sua forma. Cabe ressaltar que antes de aplicar a HT foi necessário converter as imagens para tons

de cinza e detectar as bordas por meio do Filtro de Sobel (MCREYNOLDS; BLYTHE, 2005; HAWAS; ASHOUR; GUO, 2019).

Utilizou-se o método de detecção *Hough Gradient 2-1 Hough Transform* (21HT) (YUEN et al., 1990), que executa Hough em dois estágios (ILLINGWORTH; KITTLER, 1987). No primeiro estágio, uma transformada é acumulada para encontrar as coordenadas do centro. Já no segundo estágio, um histograma de raio é construído para cada centro candidato obtido no primeiro estágio (YUEN et al., 1990). A escolha do 21HT se deu em função do método ter bom desempenho e requerer pouco espaço para armazenamento, possuindo a desvantagem de baixo desempenho na detecção de formas circulares relativamente pequenas (MARRONI, 2002), que não impactou na detecção da retina em função de seu tamanho.

Com a utilização do método 21HT foi possível encontrar a retina, que corresponde a maior circunferência presente nas imagens de fundo. Após a localização da retina, transformou-se sua circunferência no seu retângulo equivalente por meio das coordenadas x_{min} e y_{min} – referentes a posição superior esquerda na imagem – e as coordenadas x_{max} e y_{max} – referentes à posição inferior direita na imagem, conforme ilustrado na Figura 24. A área externa ao retângulo em vermelho corresponde à área de corte realizada na imagem de fundo do olho.

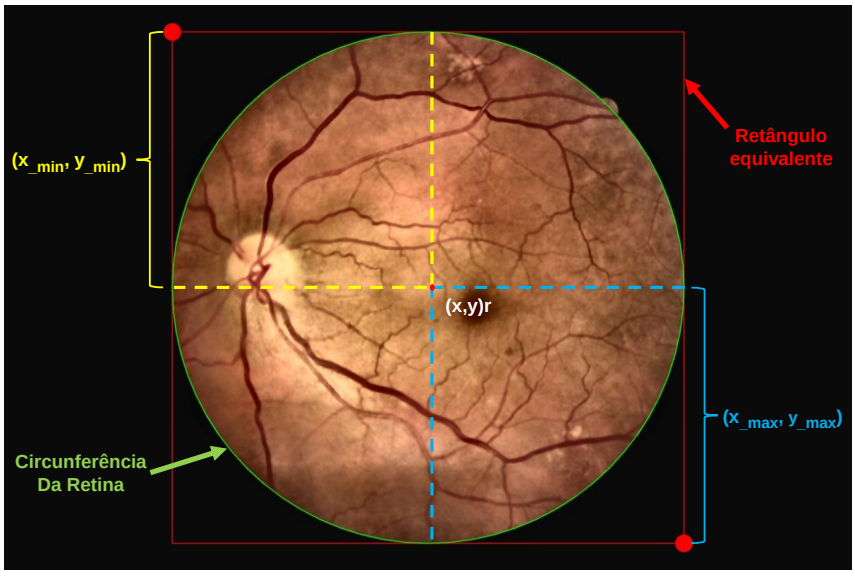


Figura 24 – Transformação da circunferência da retina em seu retângulo equivalente por meio das coordenadas x_{min} e y_{min} (posição superior esquerda), e as coordenadas x_{max} e y_{max} (posição inferior direita).

Na Figura 25, apresenta-se uma ilustração dos resultados obtidos durante a etapa de *Cropping* aplicado às imagens de fundo, em que é possível verificar a imagem resultante após a identificação do contorno da retina e o corte parcial do plano de fundo preto da imagem.

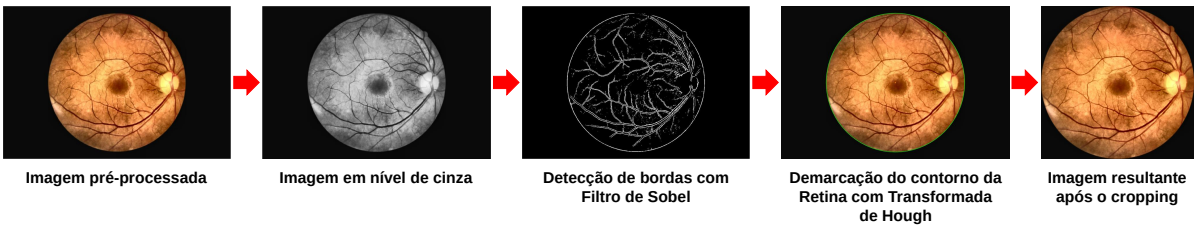


Figura 25 – *Pipeline* para realização do processo de *Cropping*. Primeiramente as imagens são pré-processadas para suavização com o Filtro de Mediana com *kernel* de tamanho 5×5. Em seguida, as imagens são convertidas para tons de cinza para a detecção das bordas com Filtro de Sobel. Após, o contorno da retina é demarcado (em verde) por meio da Transformada de Hough (21HT). Por fim, o plano de fundo preto das imagens é parcialmente recortado.

A última etapa do bloco de Pré-processamento da abordagem proposta (ver Figura 17) é a realização de *Tilling*. Redes neurais profundas que realizam a detecção de objetos em estágio único são adequadas para utilização em dispositivos de borda em virtude de seu baixo custo computacional e velocidade de inferência, principalmente quando comparadas às arquiteturas baseadas em propostas de região. Ainda assim, a detecção de objetos muito pequenos ainda permanece sendo um desafio para redes neurais profundas que realizam a detecção em estágio único (UNEL; OZKALAYCI; CIGLA, 2019).

Segundo Unel; Ozkalayci; Cigla (2019); Li; Krishna; Xu (2021), os métodos de detecção de objetos baseados em redes neurais profundas podem ser categorizados em estágio único e dois estágios. Os métodos de dois estágios são baseados em propostas de regiões, que envolve a geração de propostas de regiões seguidas de classificação, tais como os modelos R-CNN (GIRSHICK et al., 2014), Fast R-CNN (GIRSHICK, 2015), Faster R-CNN (REN et al., 2017), SPP-net (HE et al., 2014), R-FCN (DAI et al., 2016), FPN (LI et al., 2019) e Mask R-CNN (HE et al., 2020). Já os métodos de estágio único são baseados em regressão/classificação em uma estrutura unificada para detecção e classificação, tais como os modelos SSD (KONISHI et al., 2016) e YOLO (REDMON et al., 2016). Embora métodos da primeira categoria (baseados em propostas de regiões) forneçam resultados mais precisos, possuem custo computacional e tempo de inferência elevados. Já métodos da segunda categoria (baseados em regressão/classificação em uma única estrutura de rede) são mais rápidos e menos custosos computacionalmente em comparação aos modelos de dois estágios, porém possuem dificuldade em detectar objetos muito pequenos.

De modo geral, os modelos que realizam a detecção de objetos são treinados e avaliados em conjuntos de dados com grande quantidade de exemplos, tais como ImageNet (FEI-FEI; DENG; LI, 2010) e COCO (LIN et al., 2014). Estes conjuntos de dados normalmente envolvem imagens de baixa resolução (256×256, por exemplo), incluindo objetos relativamente grandes e com grande cobertura de *pixels*. Assim, modelos treinados sob estas premissas frequentemente são bem sucedidos na de-

tecção de objetos com este tipo de imagem de entrada. Entretanto, estas premissas nem sempre se refletem na detecção de objetos em imagens médicas, principalmente quando os objetos são muito pequenos e as imagens não são geradas por câmeras de alta definição (UNEL; OZKALAYCI; CIGLA, 2019).

Uma pequena cobertura de *pixels* em função da baixa resolução das imagens e/ou uma amostragem limitada de imagens afetam a capacidade de treinamento e generalização de modelos baseados em aprendizado profundo. Porém, as redes neurais profundas também possuem problemas para lidar com imagens de alta resolução em função do custo computacional necessário para processar estas imagens. Então, é necessário que estas imagens sejam redimensionadas para serem repassadas para a camada de entrada da rede neural, responsável pela leitura destas imagens. Ao realizar este processo de redimensionamento, as imagens têm sua resolução reduzida, prejudicando a extração de características das microlesões da retina. Bochkovskiy; Wang; Liao (2020) afirmam que ao otimizar uma rede neural profunda o objetivo é encontrar o equilíbrio entre a resolução da imagem de entrada da rede, a quantidade de camadas convolucionais, o número do parâmetros da rede e o número de saídas das camadas (filtros).

Nas imagens de fundo, a detecção de lesões muito pequenas (microlesões) é um desafio, como no caso de microaneurismas (vide Figura 18). Se a área da lesão não for grande suficiente, o sinal propagado nas camadas convolucionais será pequeno enquanto o treinamento do modelo é realizado, levando à dissipação de gradiente. Além disso, objetos muito pequenos são mais suscetíveis a erros de rotulagem de dados, em que a identificação da lesão pode ser prejudicada. No conjunto de dados DDR as lesões de fundo são disponibilizadas com suas anotações de caixa delimitadora (*Ground Truth*) em formato *eXtensible Markup Language* (XML). Para fins de exemplificação, no conjunto de dados DDR aproximadamente 10.978 anotações de lesões, de um total de 38.012, tem tamanho inferior a 3 *pixels*.

A abordagem que adotou-se inicialmente foi treinar os modelos de redes neurais profundas reduzindo as imagens para tamanhos como 416$\times$416, 512$\times$512 e 640$\times$640 *pixels*, levando em conta os limites máximos definidos nas arquiteturas dos modelos utilizados nos experimentos. Além disso, resoluções acima dos valores supracitados inviabilizaram a realização dos experimentos em função do elevado custo computacional e o tempo para treinamento dos modelos. Entretanto, a estratégia de reduzir o tamanho das imagens resultou em problemas oriundos das perdas de detalhes das lesões e da consequente redução da precisão da abordagem proposta.

Após a redução da resolução das imagens, algumas microlesões praticamente desapareceram, dado o tamanho diminuto destas lesões. Durante os experimentos também percebeu-se que ao utilizar imagens com resolução maior, como 640$\times$640 *pixels*, houve um aumento na precisão da detecção das lesões. Portanto, o aumento da reso-

lução das imagens utilizadas na camada de entrada da rede neural proporcionam um campo receptivo maior em torno das lesões e, com isto, um número maior de características ficam acessíveis para que as camadas convolucionais possam aprendê-las.

A solução que se adotou para evitar a necessidade de redução da resolução das imagens originais do conjunto de dados DDR, sem a utilização de GPUs com maior poder computacional e sem realizar modificações significativas na arquitetura da rede neural, foi a realização do método de *Tilling* (fatiamento), em que as imagens originais são recortadas em blocos (*tiles*).

Assim como o trabalho proposto por (UNEL; OZKALAYCI; CIGLA, 2019), implementamos o método de *Tilling* em que o tamanho dos blocos das sub-imagens resultantes são organizados de acordo o tamanho da imagem original, de acordo com a resolução e proporção da imagem original disponível no conjunto de dados. Convém destacar que o tamanho das imagens do conjunto de dados DDR, por exemplo, possuem tamanhos variados (LI et al., 2019).

Diferentemente do trabalho de (PLASTIRAS; KYRKOU; THEOCHARIDES, 2018), que apresenta um mecanismo de atenção para utilização dinâmica de *Tilling*, o método utilizado para construção dos blocos na abordagem proposta neste trabalho realiza a construção de blocos estáticos, de tamanho 2×2 (vide Figura 26), e com uma área de sobreposição entre estes blocos de 15%. As sobreposições entre os blocos são usadas para preservar os lesões ao longo dos limites destes blocos, e evitar a perda de informações devido ao fatiamento da imagem da retina.

Cada bloco corresponde a uma nova imagem para treinamento da rede neural. Com o *Tilling* o tamanho relativo das lesões é aumentado na imagem recortada em comparação com a imagem de fundo de tamanho original. Convém destacar que as lesões presentes nas imagens de fundo resultantes do fatiamento são preservadas juntamente com suas respectivas anotações (*Ground Truth*).

Nos experimentos, foi observado que com uma quantidade maior de blocos (acima de 2×2, por exemplo), as lesões que possuem tamanhos maiores (como alguns exsudatos duros e hemorragias, por exemplo), podem não caber em apenas um bloco, ou nas áreas de sobreposição dos blocos, aumentando o risco de perda da anotação destas lesões. Portanto, uma quantidade grande de blocos pode impactar na perda de informações de lesões, e por este motivo utilizamos tamanho de bloco fixo de 2×2, com o propósito minimizar o risco de perda de informações durante o treinamento da rede neural. As anotações das lesões presentes nos blocos criados são copiadas e ajustadas por meio da função de *Crop* disponibilizada pela biblioteca Albumentations (BUSLAEV et al., 2020).

Um exemplo de *Tilling* criado a partir de uma imagem de fundo de olho é ilustrado na Figura 26. Após a aplicação de *Tilling* uma quantidade maior de informações em torno das lesões das sub-imagens (blocos) foram preservadas em comparação às imagens originais que tem sua resolução reduzida (PPI) para serem utilizadas na camada de entrada da rede neural.

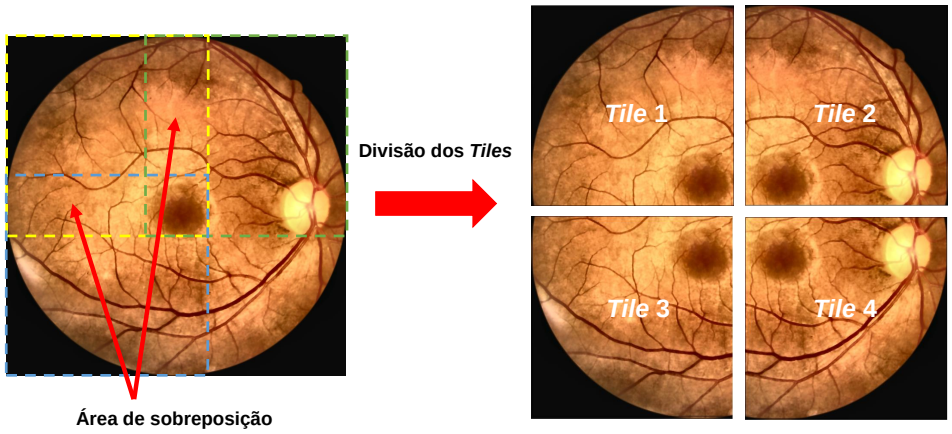


Figura 26 – À esquerda uma imagem de fundo do conjunto de dados DDR e à direita um exemplo de sub-imagens (*tiles*) criadas a partir da imagem original com *Tilling* de tamanho 2×2, e área de *overlap* de 15% entre os blocos.

Para minimizar o risco de perda de informações durante o fatiamento da imagem original definiu-se uma área de sobreposição (*overlap*) entre os blocos, conforme ilustra-se na Figura 26. Assim, não são perdidas as informações das lesões que estão presentes no local de recorte dos blocos, pois estas informações serão replicadas em diferentes blocos. Para resolver o problema de redundância de informações das lesões replicadas em diferentes blocos utilizou-se a técnica de *Non-Max Suppression* (NMS) (BODLA et al., 2017), em que é considerada apenas a caixa delimitadora com o IoU mais alto, de forma que as demais caixas delimitadoras preditas para um mesmo objeto são descartadas (MAMDOUH; KHATTAB, 2021).

Após a utilização do método de *Tilling* verificou-se um aumento na precisão da detecção das lesões, conforme apresentado nas Tabelas 12, 13, 14 e 15, em que se demonstram os resultados obtidos pela abordagem proposta nas etapas de validação e teste utilizando *Tilling* de tamanho 2×2 e sem *Tilling*.

Realizou-se *Tilling* com o propósito de reduzir a perda de detalhes das lesões em função da redução da resolução das imagens para treinamento da arquitetura da rede neural. Assim, dividiu-se as imagens em blocos para serem utilizados na camada de entrada da rede neural a fim de aumentar o campo receptivo das microlesões presentes nas imagens de fundo, uma vez que as lesões presentes nas imagens originais de alta resolução passam a ser tratadas como objetos maiores após a criação dos blocos. Mesmo com uma pequena quantidade de blocos (2×2) foi possível melhorar a precisão na detecção das lesões de fundo, conforme os resultados apresentados na

Seção 6.3.

Com o objetivo de melhorar o desempenho da abordagem proposta, principalmente na detecção das microlesões, exploraram-se técnicas de pré-processamento para obter uma melhor suavização e realce das imagens de fundo. Além disso, fez-se a remoção parcial do plano de fundo preto que ocasionava a geração de falsos positivos e realizou-se o *Tilling* das imagens originais para utilização dos blocos de imagens resultantes no treinamento da rede neural profunda. Após o pré-processamento das imagens de fundo aplicou-se uma etapa de Aumento de Dados, conforme será explicado na próxima Seção.

6.1.3 Aumento de Dados

Um empecilho frequentemente encontrado para treinamento de redes neurais profundas é a indisponibilidade de conjuntos de dados públicos com grande quantidade de objetos rotulados (DWIBEDI; MISRA; HEBERT, 2017). Este problema está relacionado com o custo para aquisição das imagens e a necessidade de especialistas para realizar as anotações dos objetos, principalmente quando se trata de imagens médicas. Ao se aumentar artificialmente o número de exemplos de um conjunto de dados reduz-se a possibilidade de subajuste do modelo, além de melhorar sua capacidade de generalização, uma vez que o modelo é treinado para predizer objetos em situações que inicialmente não estavam presentes no conjunto de dados original. As abordagens clássicas para aumento de dados em problemas de classificação de objetos incluem a aplicação de transformações geométricas básicas e transformações no espaço de cores (DVORNIK; MAIRAL; SCHMID, 2018; SHORTEN; KHOSHGOFTAAR, 2019).

Na abordagem proposta realizou-se o aumento de dados a partir das imagens e lesões rotuladas disponíveis no conjunto de dados DDR. Para cada lote de treinamento, configurou-se o modelo para passar as imagens por um carregador de dados que realiza a criação das imagens artificiais no mesmo momento em que estas são acessadas. O carregador de dados fez os seguintes tipos de aumentos: *Mosaic*, *MixUp*, *Copy-Paste* e Transformações Geométricas.

Durante o carregamento dos dados, o aumento é realizado nas imagens presentes no lote que está sendo utilizado para treinamento, ou seja, são apresentadas à rede neural diferentes versões de uma mesma imagem em cada época, porém o número de amostras permanece o mesmo. Este tipo de aumento de dados faz com que sejam criadas versões aumentadas das imagens originais a fim de proporcionar à rede neural uma maior capacidade de generalização.

Essa técnica funciona em tempo real (*on-the-fly*) e não são gerados novos exemplos antecipadamente (antes do treinamento) (NGUYEN et al., 2019; LAM et al., 2021). Assim, a cada treinamento realizado, um número aleatório de exemplos criados artificialmente são gerados e repassados para treinamento da rede neural. Cada técnica de aumento de dados é aplicada a todas as imagens do lote, com exceção do Mix-Up, que foi configurado para ser aplicado aleatoriamente em 50% das imagens do lote. A seguir, apresentam-se os detalhes sobre cada um dos métodos de aumento de dados realizados na abordagem proposta.

O primeiro método de aumento de dados aplicado foi o *Mosaic*, proposto inicialmente no trabalho de Bochkovskiy; Wang; Liao (2020), em que se combina quatro sub-imagens aleatórias de fundo para criar uma nova imagem. Convém destacar que as quatro sub-imagens que formam a imagem resultante possuem diferentes proporções. Com este método de aumento de dados foi possível que mais características das lesões em uma mesma imagem de fundo fossem extraídas. Outra vantagem foi a redução da necessidade de tamanhos de lotes muito grandes, uma vez que mais imagens foram treinadas em um mesmo lote.

Além disso, as novas imagens criadas com o *Mosaic* possibilitaram que a rede neural aprendesse características das lesões que não estavam presentes nas imagens originais, dada a aleatoriedade e assimetria das sub-imagens que compõem as novas imagens, consequentemente, aumentando a capacidade de generalização da abordagem proposta. A seguir, a Figura 27 apresenta um exemplo de aumento de dados proveniente do método *Mosaic* aplicado às imagens originais. Isto possibilitou melhorar a precisão da abordagem proposta, mesmo treinando a rede neural com tamanho de lotes menores.

Outra técnica aplicada para aumento de dados foi o *Copy-Paste*, baseada nos trabalhos de Dwibedi; Misra; Hebert (2017), Dvornik; Mairal; Schmid (2018) e Ghiasi et al. (2021), em que são geradas imagens artificiais com lesões e rótulos copiados de outras imagens de fundo do mesmo conjunto de dados. Com este método é possível criar imagens com quantidades maiores de lesões anotadas, sobre diferentes *backgrounds*. O objetivo é fazer com que a rede neural possa extrair mais características das lesões replicadas, sob diferentes contextos, aumentando a capacidade de predição destas lesões. Com a utilização do método *Copy-Paste* também é possível mitigar o problema associado à quantidade reduzida de lesões rotuladas disponíveis nos conjuntos de dados públicos de Retinopatia Diabética.

Além disso, como foi aplicado o *Copy-Paste* utilizando como base a máscara de segmentação de cada lesão disponível no próprio conjunto de dados, tanto a lesão quanto seu rótulo são copiados integralmente para a nova imagem, garantindo a integridade das informações copiadas.

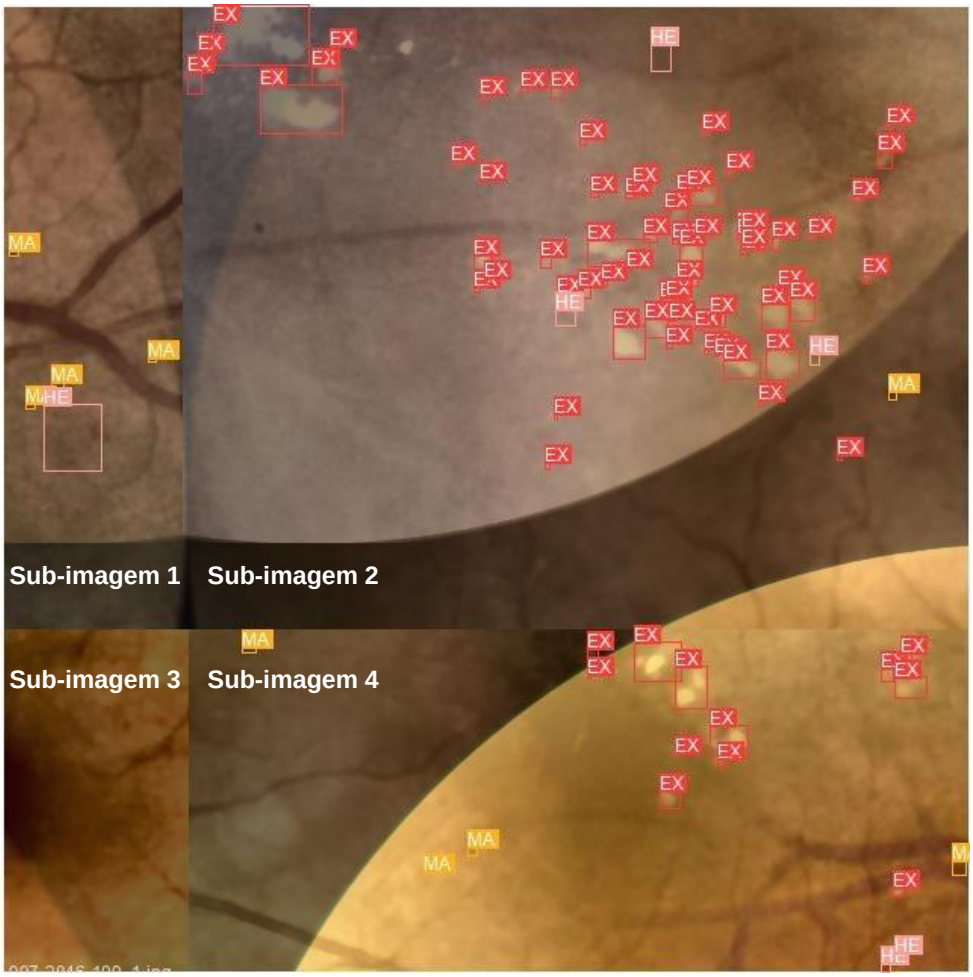


Figura 27 – Exemplo de aumento de dados obtido com o método *Mosaic*, em que quatro sub-imagens aleatórias, com diferentes proporções, são combinadas para formar uma nova imagem.

Segundo Dwibedi; Misra; Hebert (2017), em uma configuração de domínio cruzado, onde dados sintéticos são combinados com apenas 10% de dados reais, o método *Copy-Paste* supera os resultados obtidos pelos modelos treinados em apenas dados reais. Segundo os autores, a melhora no desempenho foi de mais de 21% em conjuntos de dados de *benchmark*. Ghiasi et al. (2021) também relatam uma melhoria na segmentação e detecção de instâncias do conjunto de dados COCO, tendo alcançado 49,1 de AP na detecção de máscaras e 57,3 de AP na detecção de caixas delimitadoras, uma melhoria de $+0,6$ de AP nas máscaras e $+1,5$ de AP nas caixas delimitadoras em relação ao estado da arte anterior.

O método *Copy-Paste* inicialmente foi projetado para trabalhar com a tarefa de segmentação de instância (DWIBEDI; MISRA; HEBERT, 2017) e não para detecção de objetos, uma vez que seu objetivo é copiar um objeto na íntegra e colá-lo em outra imagem. Como o YOLO trabalha com anotações na forma de caixas delimitadoras ($[class\ x, y, width, height]$) e não polígonos ($[class\ x_1, y_1, x_2, y_2, x_3, y_3, \ldots, x_n, y_n]$), comumente utilizados para a segmentação de instância, não é adequado copiar todo o

conteúdo (fundo) contido em uma caixa delimitadora juntamente com a lesão e colá-lo em uma nova imagem, pois durante o treinamento seria gerada uma grande quantidade de Falsos Positivos, prejudicando o desempenho quanto à detecção das lesões.

Para contornar este problema foram geradas as anotações das lesões na forma de polígonos a partir das máscaras de segmentação destas lesões que são disponibilizadas juntamente com as imagens de fundo do conjunto de dados DDR, conforme ilustrado na Figura 28.

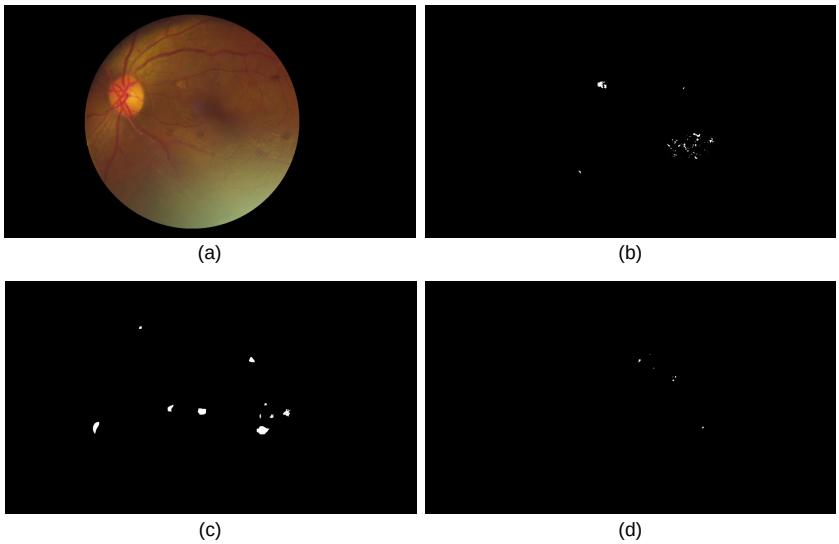


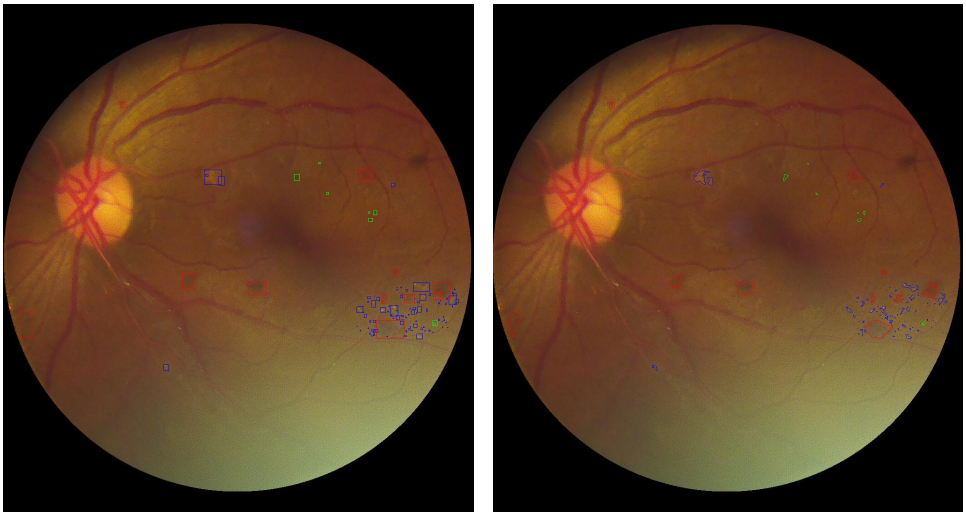
Figura 28 – Exemplo de imagem de fundo do conjunto de dados acompanhada das máscaras de segmentação das lesões presentes na imagem. Em (a), a imagem de fundo "007-3730-200.jpg" do conjunto de teste do conjunto de dados DDR; (b) máscara de segmentação dos exsudatos duros; (c) máscara de segmentação das hemorragias; e, (d) máscaras de segmentação dos microaneurismas.

Para realizar o processo de criação das anotações das lesões de fundo na forma de polígonos, primeiramente as anotações das lesões (*Ground Truth*) são importadas do conjunto de dados DDR. Em seguida, são obtidas as informações das imagens e anotações para que na sequência sejam carregados os arquivos de imagem com as máscaras binárias das lesões que estão disponibilizados no conjunto de dados. Dessa forma, com o auxílio da função `find_contours()` do OpenCV as máscaras binárias das lesões são utilizadas para capturar o contorno das lesões. Após identificar o contorno das lesões as anotações são criadas por meio da função `create_annotation_format()`. Por fim, as anotações criadas anteriormente são utilizadas para a geração de anotações no formato padrão COCO JSON por meio da função `get_coco_json_format()`.

De posse das anotações das lesões de fundo na forma de polígonos foi possível copiar apenas as lesões (excluindo o fundo em torno da lesão) e colá-las nas imagens geradas por meio da técnica *Copy-Paste*. Todavia, é importante destacar que ao treinar a rede neural é necessário que as lesões copiadas tenham anotação no

formato de caixa delimitadora para que seja possível realizar o treinamento da rede neural profunda. Para tanto, foi utilizada a função `segment2box()` do próprio YOLOv5, que converte o rótulo na forma de segmento $(xy1, xy2, \ldots)$ para um rótulo na forma de caixa (xy, xy).

A Figura 29 apresenta um exemplo de imagem de fundo do conjunto de dados DDR com as anotações das lesões de fundo, em que (a) são exemplos das anotações geradas no formato de caixas delimitadoras que são utilizadas para treinamento da abordagem proposta e (b) são exemplos das anotações geradas no formato de polígonos que são utilizadas para aplicação do método *Copy-Paste*.



(a) (b)

Figura 29 – Exemplo de imagem de fundo do conjunto de dados DDR com as anotações das lesões de fundo: (a) anotações no formato de caixas delimitadoras que são utilizadas para o treinamento da rede neural; e, (b) anotações no formato de polígonos que são utilizadas para aplicação do método *Copy-Paste*. As lesões de fundo podem ser identificas pelas cores: Exsudatos Duros (EX), em azul; Microaneurismas (MA), em verde; e, Hemorragias (HE), em vermelho.

Com o método *Copy-Paste* aumentou-se o conjunto de dados gerando dados de treinamento adicionais por meio da cópia de lesões de uma imagem para novas imagens de fundo. Basicamente, copiou-se randomicamente lesões e suas anotações de caixa delimitadora e colou-se em fundos aleatórios, assim como proposto no trabalho de Dwibedi; Misra; Hebert (2017). Na Figura 30, apresenta-se um exemplo de aumento de dados realizado com o *Copy-Paste* nas imagens de fundo do conjunto de dados DDR, em que as lesões sinalizadas foram copiadas aleatoriamente juntamente com sua caixa delimitadora para uma nova imagem.

O próximo método de aumento de dados que se utilizou foi o *MixUp* (ZHANG et al., 2018; GUO; MAO; ZHANG, 2019; CARRATINO et al., 2020), responsável por criar aleatoriamente novos exemplos por meio da combinação de imagens. A principal motivação deste método é minimizar o sobreajuste de redes neurais profundas durante

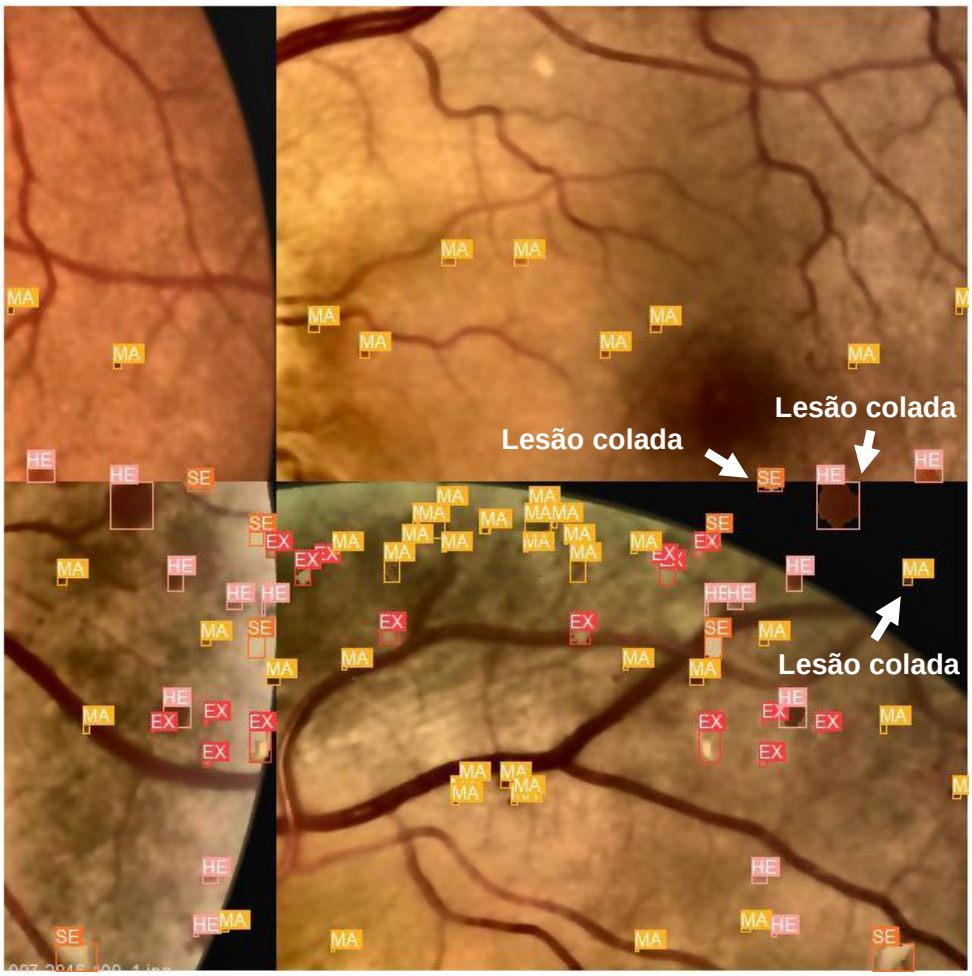


Figura 30 – Exemplo de aumento de dados realizado com o método *Copy-Paste* aplicado nas imagens de fundo. Estão sinalizadas três lesões na retina com suas caixas delimitadoras que foram copiadas aleatoriamente de outras imagens de fundo do conjunto de dados DDR e coladas aleatoriamente em uma nova imagem.

a etapa de treinamento (ZHANG et al., 2018). Apesar de sua simplicidade, *MixUp* demonstrou melhorar substancialmente a capacidade de generalização de modelos em uma ampla gama de tarefas de Visão Computacional (ZHANG et al., 2018; GUO; MAO; ZHANG, 2018).

Redes neurais profundas muito densas podem apresentar problemas de memorização e alta sensibilidade a exemplos adversários. Melhorar a capacidade de generalização e a sensibilidade a perturbações nos dados de entrada permanece um desafio (KIM et al., 2021).

A Figura 31 apresenta um exemplo de aumento de dados proveniente da utilização do *MixUp* nas imagens de fundo do conjunto de dados DDR. O propósito é combinar características de diferentes imagens, assim como as anotações de caixas delimitadoras das lesões presentes nas imagens utilizadas na combinação. Portanto, o objetivo do *MixUp* é fazer com que a rede neural não fique muito confiante sobre a relação entre características e anotações das lesões, tornando a abordagem proposta mais

sensível ao aprendizado de exemplos adversários e, consequentemente, melhorando sua capacidade de generalização. Configurou-se a abordagem proposta para aplicar o método *MixUp* aleatoriamente em 50% das imagens, pois durante os experimentos valores acima deste percentual causaram a diminuição da capacidade preditiva do modelo.

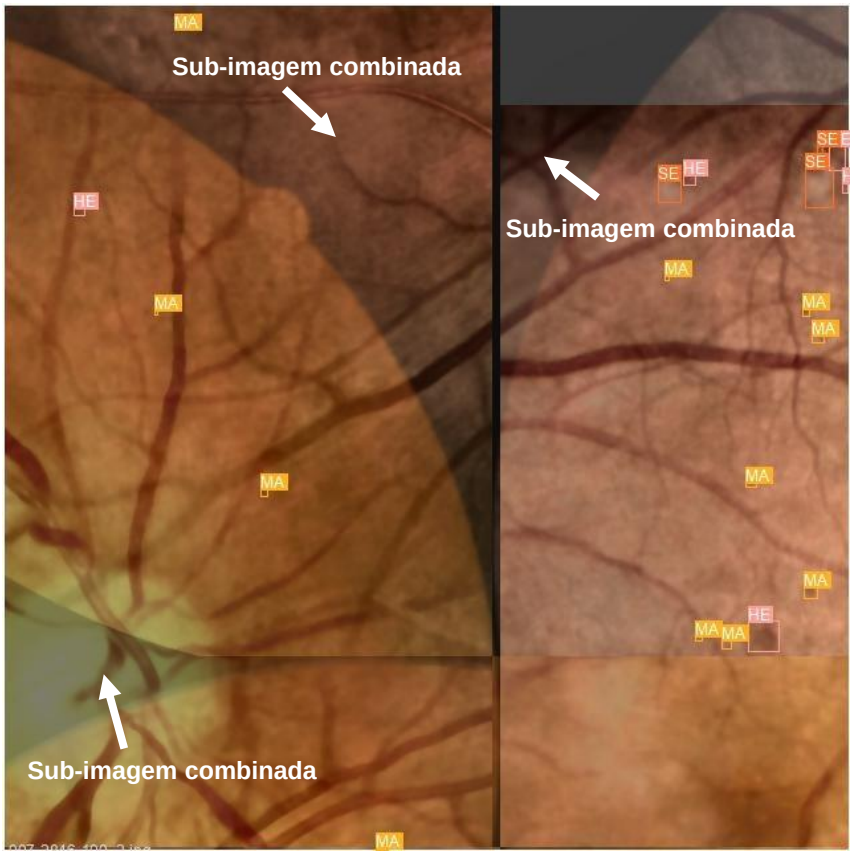


Figura 31 – Exemplo de imagem resultante da aplicação dos métodos *Mosaic*+*MixUp* em imagens de fundo do conjunto de dados DDR, em que as diferentes sub-imagens combinadas podem ser observadas conforme as sinalizações.

Também realizou-se aumento de dados por meio de transformações geométricas que foram aplicadas nas imagens originais. As transformações geométricas são facilmente implementadas e trazem boas soluções para vieses posicionais presentes nos dados de treinamento (SHORTEN; KHOSHGOFTAAR, 2019). Algumas de suas desvantagens, no entanto, incluem o custo computacional na transformação e o tempo de treinamento adicional (SHORTEN; KHOSHGOFTAAR, 2019). Além disso, algumas transformações geométricas, como translação ou cisalhamento aleatório, devem ser aplicadas com cuidado para que não sejam alterados os rótulos originais de um objeto presente na imagem.

Aplicou-se seis transformações geométricas às imagens de fundo do conjunto de dados DDR: *vertical flip up-down*, *horizontal flip left-right*, *scale*, *perspective*, *translation* e *shear*. É possível verificar na Figura 32 as transformações sobre as lesões e suas caixas delimitadoras. Durante o processo de criação das imagens artificiais, as

anotações (caixas delimitadoras) das lesões foram preservadas, de forma que não fossem perdidas as anotações originais destas lesões após a realização das transformações geométricas.

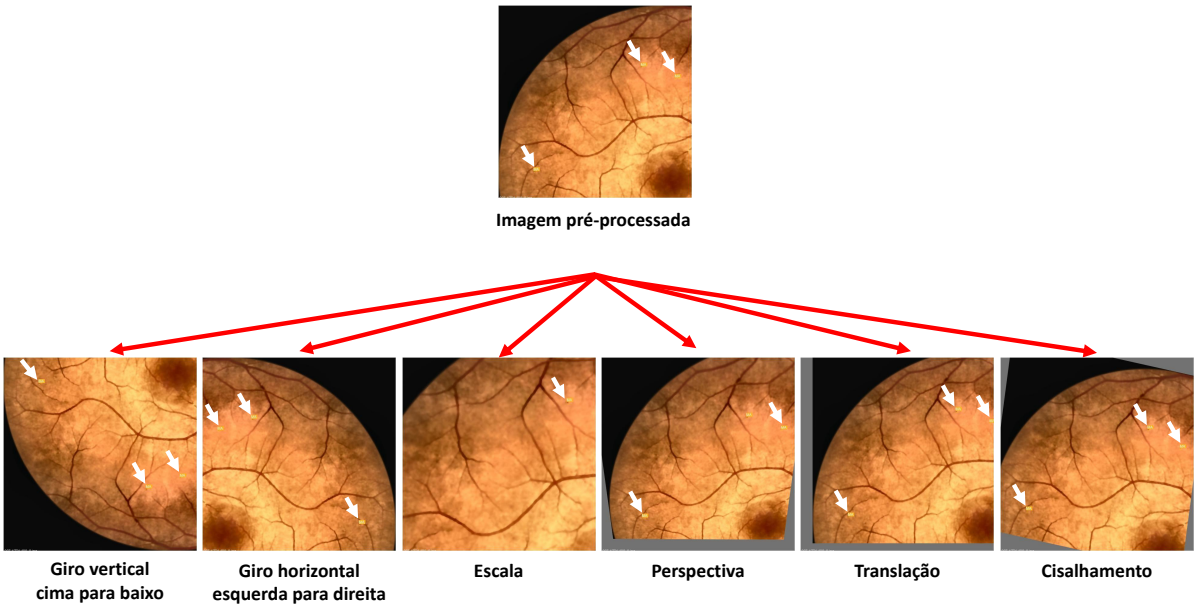


Figura 32 – Exemplo de aumento de dados sobre as imagens por meio das transformações geométricas: *Vertical flip up-down*, *Horizontal flip left-right*, *Scale*, *Perspective*, *Translation*, e *Shear*.

Na transformação geométrica *flipping* a imagem é invertida horizontal ou verticalmente. A inversão reorganiza os *pixels* enquanto protege as características da imagem. É uma das técnicas mais simples de implementar, sendo utilizada em conjuntos de dados como CIFAR-10 e ImageNet (SHORTEN; KHOSHGOFTAAR, 2019). Segundo Hao et al. (2020), o *vertical flip* vira a imagem de entrada ao longo de seu eixo x (de cima para baixo), enquanto que o *horizontal flip* vira a imagem de entrada ao longo de seu eixo y (da esquerda para a direita).

Na transformação geométrica *scaling* a imagem é dimensionada para fora (*zoom out*), ou para dentro (*zoom in*). As transformações de escala aumentam ou diminuem um determinado objeto e, como resultado, alteram os comprimentos e ângulos. O significado da transformação de escala é aumentar a coordenada xp vezes. Este requisito satisfaz $x' = xp$ e, portanto, $x = x'/p$ (SHENE, 2018).

Quando o olho humano vê uma cena, os objetos a uma certa distância parecem menores do que objetos mais próximos. Tal fenômeno é conhecido como perspectiva. A câmera funciona no mesmo princípio que a visão humana (WANG et al., 2020). O uso da transformação geométrica *perspective* é inspirada no fenômeno de perspectiva de uma câmera. Nesta transformação, o paralelismo, comprimento e ângulo não são preservados, mas sim a colinearidade e a incidência. Isso significa que as linhas retas permanecerão retas mesmo após a transformação. Também pode ser aplicado para

distorcer projetivamente uma imagem para outro plano de imagem. Por exemplo, em vez de olhar para uma cena diretamente à frente, podemos vê-la de outro ponto de vista por meio da transformação *perspective*.

Na transformação geométrica *translation* a imagem é deslocada para várias áreas ao longo do eixo x ou eixo y. Deslocar as imagens para a esquerda, direita, para cima ou para baixo pode ser uma transformação muito útil para evitar viés posicional nos dados (SHORTEN; KHOSHGOFTAAR, 2019). Uma translação desliza um objeto a uma distância fixa em uma determinada direção. O objeto original e sua translação têm a mesma forma e tamanho, e estão voltados para a mesma direção, que significa apenas mover uma imagem sem girá-la ou redimensioná-la, por exemplo.

A transformação geométrica *shearing* dá a impressão de "empurrar" um objeto geométrico em uma direção paralela a um plano de coordenadas (3d) ou um eixo de coordenadas (2d), movendo um lado da imagem e transformando-a de um formato de quadrado para um formato de trapézio (SHENE, 2018). Esta transformação é diferente da rotação porque um eixo é fixado e a imagem é esticada em um determinado ângulo chamado de ângulo de cisalhamento (CLARO et al., 2020). A Tabela 9 apresenta os parâmetros utilizados para aplicar as transformações geométricas nas imagens de fundo.

Tabela 9 – Parâmetros utilizados para aplicar as transformações geométricas nas imagens de fundo

Método	**Descrição**	**Parâmetro**
Vertical flip	Vira a entrada verticalmente em torno do eixo x	0,5
Horizontal flip	Vire a entrada verticalmente em torno do eixo y	0,5
Scale	Aplica zoom in ou zoom out na imagem transformada	(-0,5:1,5)
Perspective	Aplica perspective aleatória de quatro pontos da entrada de acordo com uma escala pré-definida	(0:0,0005)
Translation	Aplica translation na direção horizontal e vertical de acordo com uma faixa	(0,9:1,1)
Shear	Aplica shear de 15º no eixo horizontal ou vertical de acordo com uma faixa	(0,05:1,05)

Após realizar o aumento de dados teve-se que tratar também do problema relacionado ao desbalanceamento dos dados. Modelos baseados em aprendizado profundo geralmente tentam minimizar o número de erros ao classificar novos dados. Para tanto, é necessário que o custo de erros diferentes sejam iguais. Entretanto, em aplicações do mundo real, frequentemente os custos de erros diferentes são desiguais. Por exemplo, na área de diagnóstico médico, o custo de diagnosticar erroneamente

um paciente doente como saudável (Falso Negativo) pode ser muito maior do que diagnosticar acidentalmente uma pessoa saudável como doente (Falso Positivo), uma vez que o primeiro tipo de erro pode resultar na perda de uma vida (LIU et al., 2020).

É comum encontrar situações em que há desequilíbrio na quantidade de exemplos das diferentes classes de objetos que compõem um conjunto de dados. Este problema ocorre especialmente quando há uma quantidade significativamente maior de exemplos associadas a um determinado objeto. Conjuntos de dados desequilibrados são onipresentes nas tarefas de classificação e podem causar degradação do desempenho dos classificadores na previsão de amostras minoritárias (LIU et al., 2020).

Há casos em que o desbalanceamento dos dados causa vieses no treinamento dos modelos, inclusive gerando incertezas sobre os resultados obtidos (JAPKOWICZ, 2000; PROVOST, 2000). No caso das lesões de fundo do olho verificou-se um desequilíbrio da quantidade de exemplos das diferentes lesões de fundo, conforme apresenta-se na Tabela 4. Este desequilíbrio pode se tornar ainda maior após a aplicação da etapa de aumento de dados, uma vez que a quantidade de novos exemplos criados após esta etapa é aleatória, não sendo possível prever com exatidão a quantidade de novos exemplos que são gerados para cada lesão.

Mesmo que naturalmente haja uma incidência maior de algumas lesões, por vezes até em função da evolução e estágio da doença, caso o modelo seja treinado com um desbalanceamento de dados significativo, possivelmente as lesões com menor quantidade de exemplos terão uma relevância menor durante o processo de treinamento. Dessa forma, o modelo poderia de certo modo favorecer as lesões com maior probabilidade de ocorrência (classe majoritária), resultando em uma baixa taxa de reconhecimento para a classe minoritária (CASTRO; BRAGA, 2011).

Existem diversas técnicas que podem ser utilizadas para minimizar este problema como, por exemplo, a subamostragem e a sobre-amostragem do conjunto de dados (ZHOU; LIU, 2006; HE; GARCIA, 2009; FERNÁNDEZ et al., 2019). Na subamostragem são removidas amostras pertencentes à classe majoritária, de acordo com critérios projetados para reduzir o grau de desequilíbrio no conjunto de dados, ao passo que na sobre-amostragem há um aumento do número de amostras das classes minoritárias (LIU et al., 2020).

Maloof (2003) indica que um modelo aprender com conjuntos de dados desequilibrados e aprender quando os custos são desiguais ou desconhecidos são problemas que podem ser tratados de forma semelhante, sendo a aprendizagem sensível ao custo uma boa solução para ambos os casos (WEISS, 2004). Esta abordagem implica na alteração do limite de decisão, em que o processo envolve primeiramente ajustar o modelo em um conjunto de dados de treinamento e em seguida realizar as predições em um conjunto de dados de teste. As predições são realizadas na forma de probabilidades ou pontuações que são transformadas em probabilidades normali-

zadas.

Então, diferentes valores de limite são testados e os rótulos resultantes são avaliados usando uma determinada métrica de avaliação, em que o limite que atinge o melhor resultado é adotado para o modelo para fazer predições sobre novos dados. A movimentação de limite tenta mover o limite de saída para as classes de baixo custo, de forma que exemplos com custos mais altos se tornem mais difíceis de serem classificados incorretamente, sendo uma escolha relativamente boa para o treinamento de redes neurais sensíveis ao custo (ZHOU; LIU, 2006).

Para equilibrar o número de exemplos de cada lesão, durante o treinamento utilizou-se o método *Threshold-Moving* (ZHOU; LIU, 2006; BUDA; MAKI; MAZUROWSKI, 2018; ZHANG; GWEON; PROVOST, 2020) por meio do parâmetro `image-weights`, em que as amostras de imagens do conjunto de treinamento são ponderadas por seu *mean Average Precision* (mAP) inverso do teste da época anterior. Ou seja, ao contrário de amostrar as imagens uniformemente durante o treinamento, como é realizado em um treinamento convencional, a amostragem durante o treinamento toma por base as imagens ponderadas com base no resultado obtido por uma determinada métrica de avaliação calculada no teste da época anterior do treinamento. Este método move o limite de decisão de modo que os exemplos de classe minoritária sejam mais fáceis de serem previstos corretamente (HE; GARCIA, 2009; FERNÁNDEZ et al., 2019).

Utilizou-se o método *Threshold-Moving* com o intuito de minimizar o desbalanceamento do conjunto de dados e reduzir a possibilidade de vieses durante a classificação em função da presença de classes com quantidade significativamente maior de exemplos. Em função da necessidade de grande quantidade de exemplos para a obtenção de resultados mais precisos optou-se por não utilizar a técnica de subamostragem das classes majoritárias. A sobre-amostragem das classes minoritárias também foi descartada pois o aumento do número de exemplos destas classes com o objetivo de equipará-las às classes majoritárias não refletiria a incidência natural das lesões de fundo.

As técnicas de sobre-amostragem e sub-amostragem modificam a distribuição dos dados de treinamento de modo que os custos dos exemplos são transmitidos explicitamente (ZHOU; LIU, 2006). Além disso, como o aumento de dados gera uma quantidade aleatória de exemplos artificiais, não há como prever com exatidão o número total de exemplos para cada lesão, não sendo portanto possível definir um valor exato para sub-amostrar ou sobre-amostrar um determinado tipo de lesão.

Utilizou-se um método para minimizar o problema de desbalanceamento da quantidade de exemplos das lesões a fim de evitar um possível sobreajuste do modelo associado à classificação errônea de lesões na classe majoritária, com o objetivo de aprimorar a capacidade de generalização da abordagem proposta, tomando cuidado

para que o método utilizado não descaracterizasse completamente a incidência naturalmente maior de algumas lesões, com a exclusão ou adição arbitrária de exemplos em uma determinada classe de lesões. Após a etapa de aumento e balanceamento dos dados, treinou-se a abordagem proposta para realizar a detecção das lesões de fundo do olho. Na próxima Seção será apresentada a arquitetura da rede neural profunda utilizada na abordagem proposta.

6.2 Arquitetura da Rede Neural Profunda

Após as etapas de pré-processamento e aumento de dados descritas anteriormente, dividiu-se o conjunto de imagens do conjunto de dados DDR em conjunto de treinamento (50%), conjunto de validação (20%) e conjunto de teste (30%), mesma proporção realizada no trabalho de Li et al. (2019). As imagens de um conjunto não estão presentes nos demais, a fim de evitar enviesamento durante a avaliação da abordagem proposta. Utilizou-se uma etapa de validação para realizar o ajuste fino dos hiperparâmetros da arquitetura e uma etapa de teste para aferir a capacidade de generalização da rede neural. A fim de validar a abordagem proposta também utilizou-se o conjunto de dados público de Retinopatia Diabética IDRiD (PORWAL et al., 2020), no qual se aplicou o mesmo método de particionamento adotado para o conjunto de dados DDR.

Utilizou-se como base da abordagem proposta a arquitetura do modelo YOLOv5 (GE et al., 2021; ALYOUBI; ABULKHAIR; SHALASH, 2021; COUTURIER et al., 2021; QI et al., 2021; ZHU et al., 2021; RAHMAN; AZAD; HASAN, 2021; ZHENG; ZHAO; LI, 2021; XU et al., 2021; ZHU et al., 2021). Este modelo possui atualmente quatro versões: s (*small*), m (*medium*), l (*large*) e x (*extra-large*). Cada versão possui diferentes características. No entanto, a principal diferença está no valor atribuído aos multiplicadores de escala da largura e profundidade da rede (RAHMAN; AZAD; HASAN, 2021), inspirada no modelo EfficientDet (TAN; PANG; LE, 2020). Foram realizados experimentos com as diferentes versões do YOLOv5 e optou-se pela versão s para servir como arquitetura de base para a abordagem proposta. As quantidades de multiplicadores de profundidade (*depth*) e de escala (*width*) de núcleos convolucionais do modelo adotado são 0,33 e 0,50, respectivamente. Segundo Iyer et al. (2021), o YOLOv5s atinge precisão equivalente ao YOLOv3, porém com desempenho superior na realização de inferências em tempo real, a um custo computacional inferior.

As Redes Neurais Convolucionais apresentam dezenas ou centenas de milhões de parâmetros, o que impõe uma sobrecarga de computação e memória considerável. Isto pode acabar ocasionando problemas para realizar o treinamento, otimização e eficiência de memória (HASANPOUR et al., 2018). Assim, a utilização de um modelo com uma quantidade menor de parâmetros possibilitou que fossem utilizados menos

recursos de *hardware*, viabilizando portanto a detecção das lesões de fundo por meio de GPUs de baixo custo sem, no entanto, impactar na precisão da abordagem proposta. A estrutura da rede neural profunda utilizada em nossa abordagem possui um total de 283 camadas, 7,2 milhões de parâmetros e 17,1 GFLOPs. Segundo Yu et al. (2021), GFLOPs representa a 1 bilhão de FLOPs (operações matemáticas de ponto flutuante).

Outra vantagem da utilização do modelo YOLOv5 como base desta abordagem está na possibilidade de integração e portabilidade com diferentes tipos de projetos, inclusive para dispositivos móveis, por exemplo. Esta característica está associada ao fato deste modelo ter sido implementado nativamente em PyTorch. Outra característica importante do YOLOv5 é a possibilidade de automatizar o processo de geração das âncoras das caixas delimitadoras dos objetos. Com isso, o cálculo dos tamanhos das âncoras é realizado com base nas lesões de fundo do conjunto de dados utilizado nos experimentos.

Utilizou-se o *Backbone* da rede como extrator de características pré-treinado em um conjunto de dados de classificação de imagens, útil para detectar objetos nas últimas camadas da rede. O *Backbone* usado nos experimentos é composto por uma estrutura formada por um *Cross Stage Partial Network* (CSP) e uma rede neural convolucional Darknet-53.

O modelo CSP (WANG et al., 2020) é baseado na arquitetura da rede neural profunda DenseNet (HUANG et al., 2017), que foi projetada para reduzir o problema de dissipação de gradiente, reforçar a propagação e reutilização de características e reduzir o número de parâmetros da rede. Portanto, o CSP minimiza problemas de gradiente que ocorrem em *backbones* que utilizam redes convolucionais maiores, mesmo com uma quantidade menor de parâmetros, que o torna uma boa alternativa para modelos baseados em arquiteturas da família YOLO, em que a velocidade de inferência e o tamanho das redes neurais profundas são relevantes.

Redes neurais aplicadas a problemas de classificação de imagens e detecção de objetos apresentam resultados no estado da arte, especialmente quando se tornam mais profundas (HE et al., 2016; XIE et al., 2017; HUANG et al., 2017) e mais largas (ZAGORUYKO; KOMODAKIS, 2017). Porém, ao desenvolver arquiteturas de redes neurais mais profundas, por exemplo, geralmente há um número maior de cálculos e, por conseguinte, um custo computacional maior que tornam tarefas como a detecção de objetos inacessíveis para a maioria das pessoas, uma vez que problemas do mundo real exigem um tempo de inferência rápido e em dispositivos com baixo poder computacional (WANG et al., 2020).

Neste contexto, o modelo CSP proposto por Wang et al. (2020) pode ser aplicado em diferentes arquiteturas, tais como ResNet (HE et al., 2016), ResNeXt (XIE et al., 2017), DenseNet (HUANG et al., 2017), YOLOv4 (BOCHKOVSKIY; WANG; LIAO, 2020) e YOLOv5 (XU et al., 2021; ZHU et al., 2021), pois não apenas produz uma redução no custo computacional e o uso de memória dessas redes, mas também traz benefícios como a melhoria na velocidade de inferência e o aumento da precisão. Estes objetivos são alcançados mediante o particionando do mapa de características da camada base em duas partes. Em seguida, as partes são mescladas por meio de uma hierarquia de estágios cruzados, cuja ideia principal é fazer com que o gradiente se propague por diferentes caminhos da rede.

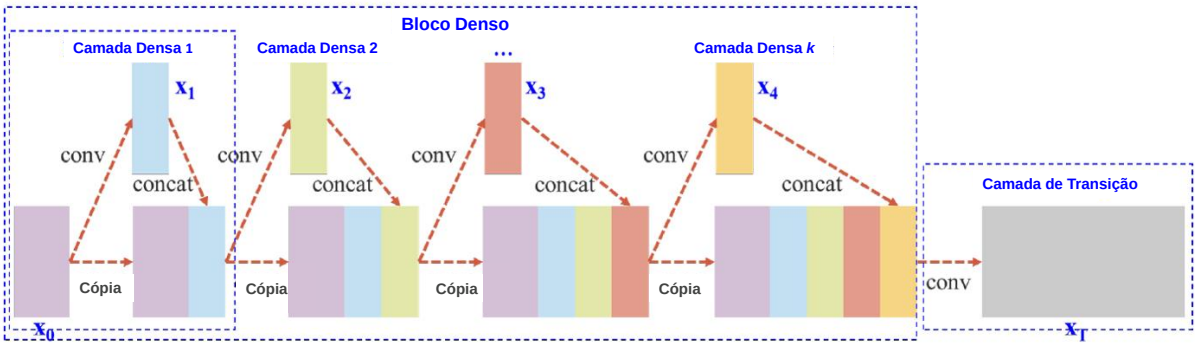


Figura 33 – Bloco Denso em uma arquitetura de rede neural profunda DenseNet (WANG et al., 2020).

A Figura 33 apresenta um bloco denso em uma arquitetura de rede neural profunda DenseNet (HUANG et al., 2017), em que a saída da i-ésima camada densa será concatenada com a entrada da i-ésima camada densa. Este resultado concatenado torna-se a entrada da $i+1$-ésima camada densa, conforme a Equação 19 (WANG et al., 2020):

$$\begin{aligned} x_1 &= w_1 * x_0 \\ x_2 &= w_2 * [x_0, x_1] \\ &\vdots \\ x_k &= w_k * [x_0, x_1, \ldots, x_{k-1}] \end{aligned} \tag{19}$$

Se for utilizado um algoritmo de retro-propagação para atualizar os pesos, as equações de atualização de peso podem ser escritas como:

$$
\begin{aligned}
w_1' &= f(w_1, g_0) \\
w_2' &= f(w_2, g_0, g_1) \\
w_3' &= f(w_3, g_0, g_1, g_2) \\
w_k' &= f(w_k, g_0, g_1, g_2, \ldots, g_{k-1})
\end{aligned}
\tag{20}
$$

Com a hierarquia de estágios cruzados da CSP, uma grande quantidade de informações de gradiente é reutilizada para atualizar pesos de diferentes camadas densas. Este processo resultará em diferentes camadas densas aprendendo repetidamente as informações de gradiente copiadas.

A Figura 34 apresenta um exemplo de CSP, em que é possível observar que o mapa de características da camada de base é separado em duas partes, uma parte passará por um bloco denso e uma camada de transição, a outra parte é então combinada com o mapa de características transmitido para o próximo estágio.

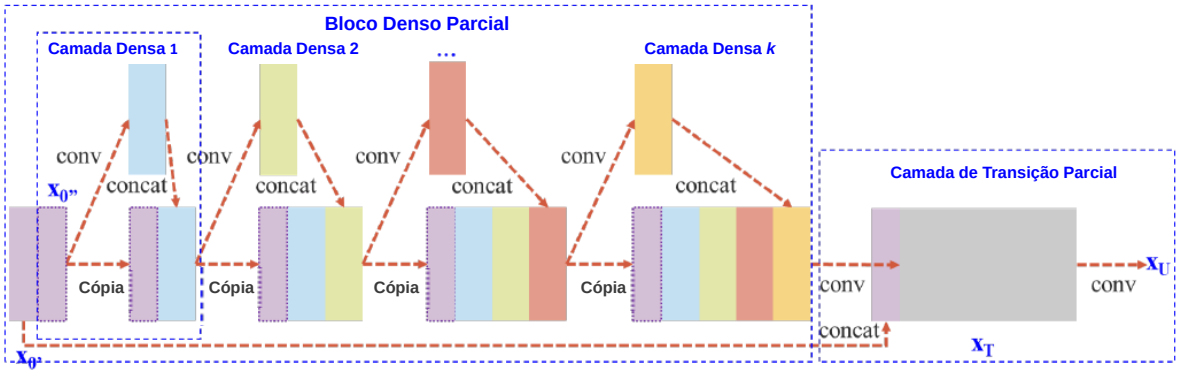


Figura 34 – *Cross Stage Partial DenseNet* (CSPDenseNet) (WANG et al., 2020).

As equações de passagem *feed-forward* e atualização de peso da CSPDenseNet tornam-se (WANG et al., 2020):

$$
\begin{aligned}
x_k &= {w_k}^*[x_{0''}, x_1, \ldots, x_{k-1}] \\
x_T &= {w_T}^*[x_{0''}, x_1, \ldots, x_k] \\
x_U &= {w_U}^*[x_{0'}, x_T]
\end{aligned}
\tag{21}
$$

$$
\begin{aligned}
{w_k}' &= f(w_k, g_{0''}, g_1, g_2, \ldots, g_{k-1}) \\
{w_T}' &= f(w_T, g_{0''}, g_1, g_2, \ldots, g_k) \\
{w_U}' &= f(w_U, g_{0'}, g_T]
\end{aligned}
$$

onde os gradientes provenientes das camadas densas são integrados separadamente. O mapa de características que não passou pelas camadas densas também é integrado separadamente. Quanto às informações para atualização de pesos, ambos os lados não contêm informações duplicadas de gradiente. Portanto, o CSPDenseNet preserva a vantagem de reutilização de características da DenseNet, porém evita uma quantidade excessiva de informações duplicadas ao truncar o fluxo de gradiente. De acordo com Wang et al. (2020), dividindo as entradas apenas metade dos mapas de características passam pelo bloco denso parcial, portanto, o cálculo também diminui em 50%. O pseudocódigo do CSPDenseNet é descrito no Algoritmo 2 (WANG et al., 2020).

Algoritmo 2 Pseudocódigo do CSPDenseNet

1: A saída do bloco denso anterior é dividida pela metade ▷ `Exemplo: o mapa de características de 60 imagens será dividido em 30, 30.`
2: Um conjunto da saída passará pelo bloco denso parcial ▷ `Parcial, pois apenas uma parte de todo o mapa de características é enviada através do bloco denso.`
3: O segundo conjunto de mapas de características está diretamente vinculado ao final deste estágio ▷ `Ou seja, após o bloco denso.`
4: A saída das camadas densas passará por uma camada de transição parcial (redimensionamento e dimensionamento) e, então, será concatenada com o segundo conjunto de mapas de características. ▷ `Essa saída concatenada passará novamente por uma camada de transição parcial.`

Já a rede neural convolucional Darknet-53 foi inicialmente utilizada como *Backbone* do modelo YOLOv3 (REDMON; FARHADI, 2018), em substituição ao seu antecessor Darknet-19 (REDMON et al., 2016), pois inclui o uso de conexões residuais, bem como mais camadas, uma vez que possui 53 camadas de profundidade (REDMON; FARHADI, 2018). Sua arquitetura é construída com camadas consecutivas de convo-

lução 3×3 e 1×1 seguidas por uma conexão de salto, que auxilia as ativações a se propagarem através de camadas mais profundas sem a dissipação de gradiente.

O diagrama de blocos da arquitetura da rede neural que compõe a abordagem, responsável pela detecção das lesões de fundo, é ilustrado na Figura 35. Como é possível observar a arquitetura da rede é baseada na estrutura do YOLOv5 versão s (IYER et al., 2021; XU et al., 2021; ZHU et al., 2021) e está dividida em três blocos principais: *Backbone*, *Neck* e *Head*.

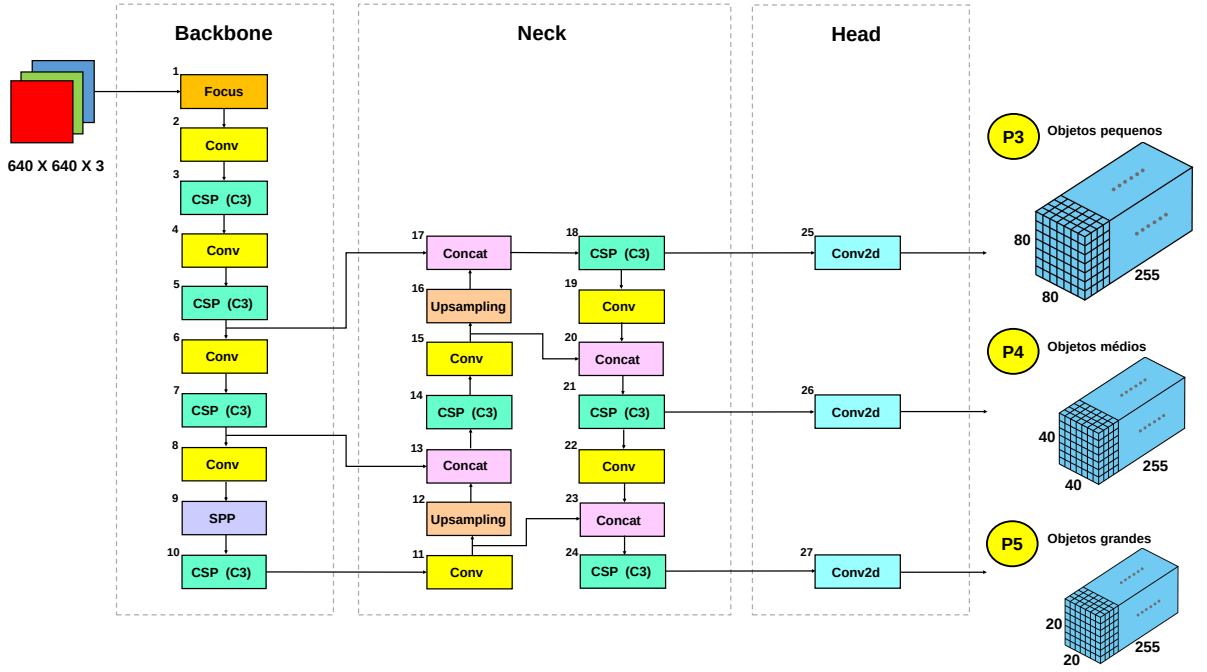


Figura 35 – Diagrama de blocos da arquitetura da rede neural que compõe a abordagem para a detecção das lesões de fundo. A estrutura é dividida em três blocos principais: *Backbone*, *Neck* e *Head*. A entrada da rede recebe imagens de tamanho 640×640×3 e a saída é composta por três cabeças de detecção: a camada P3, responsável pela detecção de objetos pequenos; a camada P4, responsável pela detecção de objetos médios; e, por fim, a camada P5, responsável pela detecção de objetos grandes.

O tamanho da camada de entrada da rede é de 640×640×3, em que os dois primeiros valores correspondem à altura e largura em *pixels* e o terceiro valor corresponde à quantidade de canais da imagem de entrada. O conjunto de dados DDR disponibiliza as anotações de caixa delimitadora das lesões de fundo para fins de treinamento de redes neurais profundas. Foram utilizadas estas caixas delimitadoras para realizar o cálculo e ajuste dos tamanhos das âncoras. Em arquiteturas da família YOLO normalmente são configurados os tamanhos das âncoras antes de realizar o treinamento do modelo.

No treinamento da abordagem são produzidas caixas delimitadoras com as predições baseadas nos comprimentos das âncoras ajustadas. Em seguida é realizada uma comparação da caixa delimitadora do objeto detectado com a caixa delimitadora anotada do objeto (*Ground Truth*). O resultado desta comparação é então utilizado

na atualização dos pesos da rede neural durante a etapa de treinamento. Portanto, a definição do tamanho das âncoras é importante, principalmente quando o treinamento da rede neural é realizado sobre objetos com tamanhos diferentes das dimensões padrão de âncoras, que normalmente são calculadas com base em objetos de conjuntos de dados como o COCO, por exemplo.

O repositório do modelo YOLOv5 dispõe de uma função que realiza o cálculo adaptativo das âncoras, em que no momento do treinamento da rede neural é possível habilitar a opção "`auto-anchor`" para que sejam calculados os melhores valores para as caixas de ancoragem de forma automática. Foi utilizada esta função antes do treinamento da abordagem para garantir que as âncoras fossem ajustadas de acordo com os tamanhos das lesões de fundo presentes no conjunto de dados utilizado no experimento. O recurso de `auto-anchor` funciona por meio de um algoritmo de aprendizado não supervisionado K-means e um Algoritmo Genético (AG). Em suma, uma otimização com um AG é executada nas âncoras após uma varredura realizada com K-means (LI et al., 2021). Utilizou-se $k = 9$ para o número de *clusters*, que corresponde à quantidade de âncoras das cabeças de detecção que constituem o *Head* da rede neural. A função de distância no agrupamento é obtida pela Equação 22 (LI et al., 2021):

$$d(box, centroid) = 1 - IoU(box, centroid) \tag{22}$$

onde: $centroid$ representa o centro do *cluster*, box (anchor box) representa a amostra, e $IoU(box, centroid)$ é a intersecção sobre a união da box do centroid e a box do *cluster* (LI et al., 2021). Utilizou-se $k = 9$ para obter a relação entre o número de centros de agrupamento k e o *average* da interseção sobre união. Os tamanhos das âncoras ajustadas para cada cabeça de detecção são apresentados na Tabela 11. A função *fitness* do AG é calculada com base na função de perda de regressão de caixa delimitadora e o número de gerações definido para o AG realizar a otimização das âncoras foi $gen = 1000$.

A estrutura do *Backbone* da rede neural inicia com o módulo *Focus*, responsável por realizar uma operação de fatiamento. Para fins de exemplo, na Figura 36(b) ilustra-se uma imagem de entrada com tamanho 4×4×3 e seu fatiamento para um mapa de características de 2×2×12. No caso da estrutura da rede neural da abordagem, quando uma imagem de tamanho 640×640×3 é inserida no módulo *Focus*, é realizada uma operação de fatiamento desta imagem para a geração de um mapa de características de tamanho 320×320×64, conforme ilustrado na Figura 36(a), cujo objetivo é prover uma melhor extração de características durante o *downsampling* da imagem (ZHU et al., 2021).

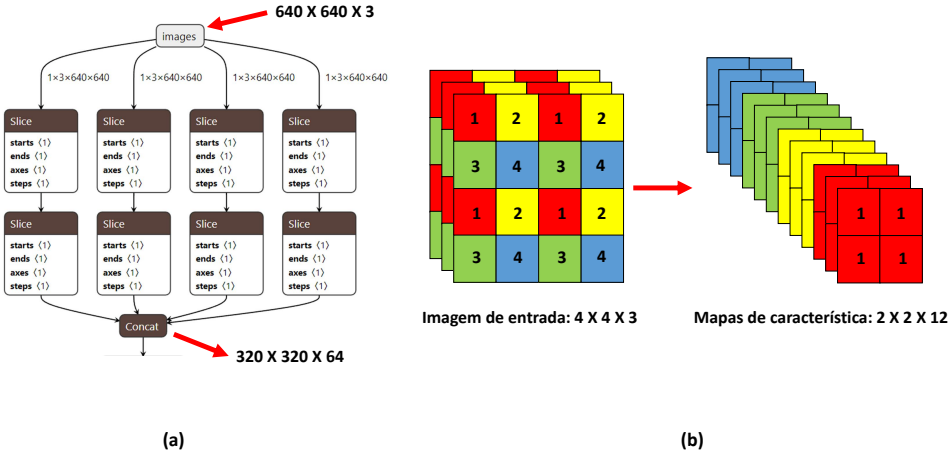


Figura 36 – Estrutura do sub-bloco *Focus* do *Backbone* da rede. Em (a), uma operação de fatiamento é realizada em um imagem de tamanho 640×640×3 para a geração de um mapa de características de tamanho 320×320×64 para promover uma extração de características melhorada durante o *downsampling* da imagem; e, (b), uma imagem de entrada com tamanho 4×4×3 e seu fatiamento para um mapa de características de tamanho 2×2×12.

Ainda no *Backbone*, os módulos *Conv* são compostos por uma convolução 2d, seguida por uma normalização de lote. Esta normalização é utilizada para padronizar as entradas e estabilizar o processo de aprendizagem, uma vez que a distribuição das entradas de cada camada muda durante o treinamento. Além disso, a normalização de lote reduz o número de ciclos de treinamento necessários para treinar redes profundas, fornecendo uma regularização e reduzindo o erro de generalização (IOFFE; SZEGEDY, 2015).

Após a normalização de lote é aplicada a função de ativação *Sigmoid Linear Unit* (SiLU) (ELFWING; UCHIBE; DOYA, 2017), derivada da função *Rectified Linear* (ReLU) (AGARAP, 2019). A função SiLU é calculada pela função sigmoide multiplicada por sua entrada e, segundo Elfwing; Uchibe; Doya (2017), apresenta melhores resultados que a função ReLU. Utilizou-se a função SiLU em toda a estrutura da rede neural com o intuito de simplificar a arquitetura, uma vez que é utilizado apenas um tipo de função de ativação. A função SiLU ak é obtida conforme a Equação 23 (NWANKPA et al., 2018):

$$a_k(S) = z_k \alpha(z_k) \tag{23}$$

onde, S é o vetor de entrada e z_k é a entrada para unidades ocultas k. A entrada para as camadas ocultas é fornecida pela Equação 24:

$$z_k = \sum_i w_{ik} S_i + b_k \tag{24}$$

onde, b_k é o bias e w_{ik} é o peso que conecta as unidades ocultas k respectivamente.

Na arquitetura da rede é utilizado o módulo CSP (C3) tanto no *Backbone* quanto no *Neck* da rede. Estas CSPs são utilizadas para conectar as camadas frontais e posteriores da rede, visando melhorar a velocidade de inferência do modelo sem comprometer sua precisão. Também possibilitam uma melhor integração das diferentes partes que compõem a rede neural, além de uma redução do tamanho do modelo (ZHU et al., 2021). Estes módulos C3 têm em sua estrutura três módulos *Conv* e um módulo *Bottleneck* (HE et al., 2016). O módulo *Bottleneck* (Figura 37), é constituído de dois módulos Conv seguidos de uma operação de adição (*add*), responsável por adicionar tensores sem expandir a dimensão da imagem.

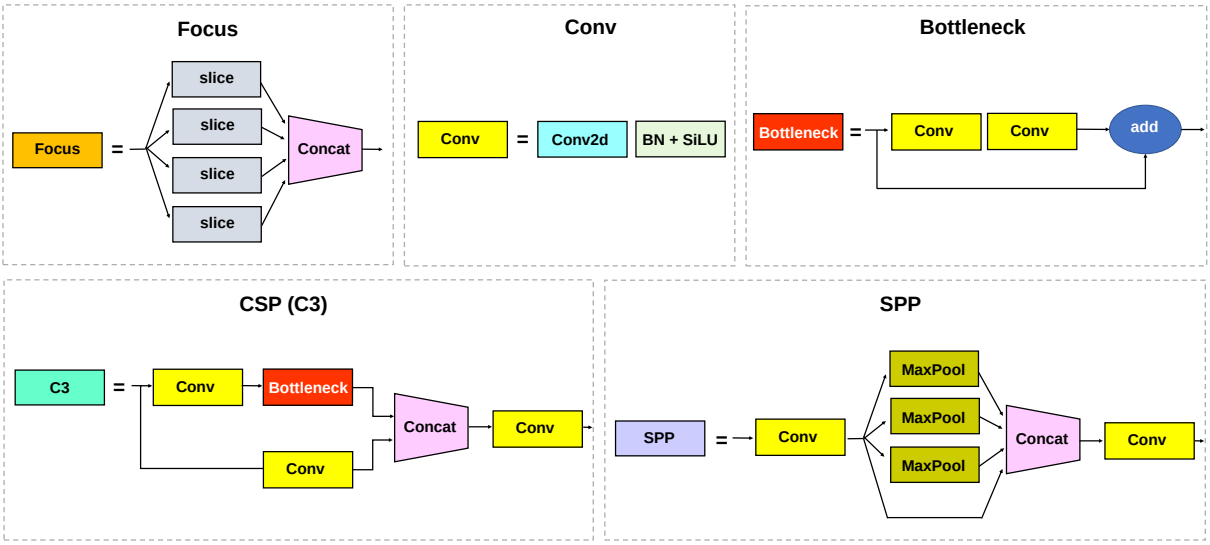


Figura 37 – Sub-blocos que compõem os blocos principais da arquitetura de rede neural da abordagem, dentre os quais o módulos *Focus*, Conv, *Bottleneck*, CSP (C3) e SPP.

Um bloco *BottleNeck* (Figura 38b) é semelhante a um bloco residual comum que funciona como uma conexão de salto (atalho). Inicialmente proposto no trabalho de He et al. (2016), os autores empilharam vários desses blocos residuais um após o outro para formar uma rede neural residual profunda. A Figura 38a apresenta um exemplo de bloco residual comum de uma *Residual Network* (ResNet) (HE et al., 2016). É possível observar que o bloco residual permite que as camadas continuem a receber os valores resultantes da função de ativação $F(x)$ da camada anterior e também os valores da entrada x destas funções.

O uso de blocos residuais permite uma redução da dissipação de gradiente que ocorre no treinamento de redes neurais substancialmente mais profundas (HE et al., 2016). *BottleNeck* é uma variante de blocos residuais, entretanto, utiliza convoluções com filtro de tamanho 1×1 para criar um gargalo. O bloco *BottleNeck* aplica uma convolução 1×1 para reduzir as dimensões, em seguida uma convolução 3×3 e, por fim, uma convolução 1×1 para aumentar (restaurar) as dimensões. Esta operação é mais rápida que aplicar três convoluções com filtro de tamanho 3×3. Segundo He et al. (2016), as convoluções 1×1 são utilizadas para criar um gargalo com o intuito de

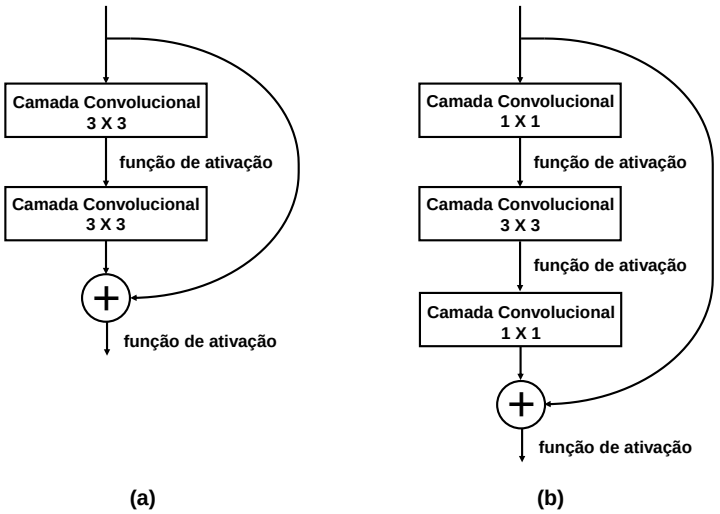


Figura 38 – Exemplo de bloco residual comum (a) versus bloco *Bottleneck* (b). O bloco *Bottle-Neck* é uma variante do bloco residual. Primeiro, uma convolução 1×1 reduz a as dimensões e depois outra convolução 1×1 restaura as dimensões com o propósito de criar um gargalo e diminuir a quantidade de parâmetros e multiplicações de matrizes. Fonte: Adaptado de He et al. (2016).

reduzir a quantidade de parâmetros e multiplicações de matrizes, fazendo blocos residuais tão finos quanto possível para aumentar a profundidade e ter menos parâmetros. O *Backbone* da abordagem é composto por 4 módulos CSP (C3).

Em cada módulo C3, após o módulo *Bottleneck*, há um módulo de concatenação (*Concat*) para que as características que foram divididas no início do bloco C3 sejam reagrupadas, expandindo as dimensões dos tensores. Por exemplo, a concatenação de um mapa de características de tamanho $2\times26\times512$ com outro de tamanho $26\times26\times512$ tem como resultado um mapa de características de tamanho $26\times26\times1024$. O fluxo dos diversos módulos que compõem o *Backbone*, bem como a constituição de cada um destes módulos, pode ser observado nas Figuras 35 e 37.

Outro componente do *Backbone* da abordagem é o módulo SPP (*Spatial Pyramid Pooling*) (HE et al., 2014). CNNs são usadas para extração de características de imagens, seguidas por camadas totalmente conectadas (FC) para realizar a classificação. Como a operação de convolução é aplicada em uma janela deslizante, *a priori* seria possível que a entrada da rede tivesse um tamanho variável, resultando em saída de tamanho variável. Entretanto, como as CNNs são seguidas por camadas FCs que só aceitam entradas de tamanho fixo, as CNNs se tornam incapazes de aceitar entradas com tamanhos variáveis.

Assim, as imagens primeiramente são redimensionadas para um determinado tamanho antes de serem inseridas na CNN. Este processo pode gerar problemas como distorção da imagem e também a perda de resolução e informações. Portanto, a restrição de imagens de tamanho fixo para a camada de entrada da rede neural não se deve às camadas convolucionais e sim às camadas FCs, que devem ter uma entrada

vetorial de comprimento fixo. Para contornar este problema, He et al. (2014) substituíram a última camada de *pooling* (camada antes da FC) por uma camada SPP. Normalmente, uma CNN tem uma única camada de *pooling* ou nenhum *pool* antes das camadas FC, mas com a SPP é possível introduzir vários *pools* de escala variável que são concatenados para formar um vetor de 1 dimensão para a camada FC, conforme ilustrado na Figura 39.

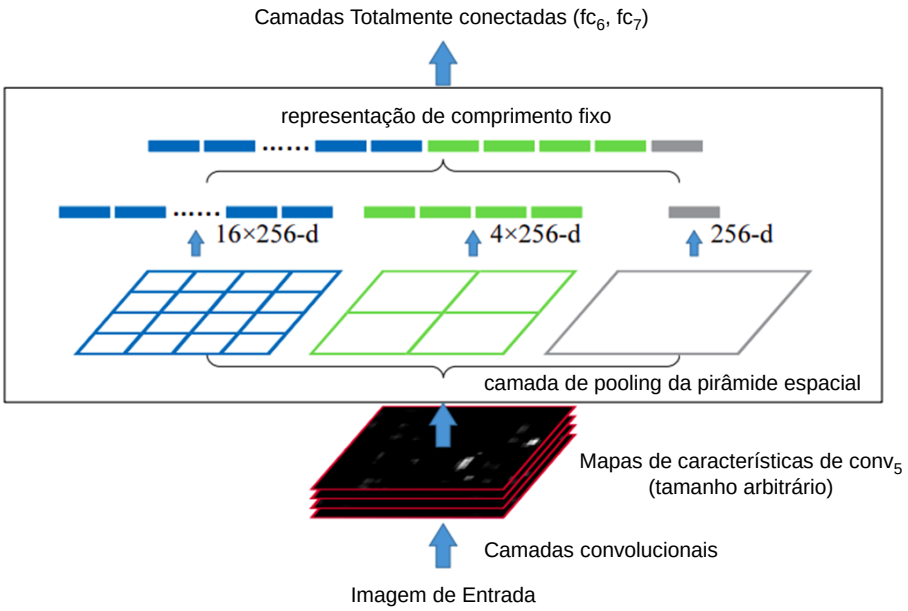


Figura 39 – Exemplo de estrutura de rede com uma *spatial pyramid pooling layer* com 3 escalas/pirâmides, sendo conv5 a última camada convolucional e 256 o número de filtros desta camada. Fonte: He et al. (2014).

A SPP mantém as informações espaciais em compartimentos espaciais locais. O número de caixas e seu tamanho são fixos. Em cada categoria espacial as respostas de cada filtro são agrupadas. Na Figura 39, é realizado um *pooling* de três níveis, em que (HE et al., 2014):

- O mapa de características de saída tem 256 filtros e possui tamanho de acordo com o tamanho de entrada.
- Na primeira camada de *pooling*, a saída tem um único compartimento e cobre uma imagem completa, semelhante à operação de *pool* global. A saída desse *pool* é de 256 dimensões.
- No segundo agrupamento, o mapa de características é agrupado para ter 4 compartimentos, resultando em uma saída de tamanho 4×256.
- No terceiro agrupamento, o mapa de características é agrupado para ter 16 compartimentos, resultando em uma saída de tamanho 16×256.

Dessa forma, a saída de todas as camadas do agrupamento é achatada e concatenada para fornecer uma saída de dimensão fixa, independentemente do tamanho da entrada. Assim como em He et al. (2014), foi utilizado o método de agrupamento *MaxPool* com agrupamentos de tamanho igual a 1×1, 5×5, 9×9 e 13×13, seguido da operação *Concat* para concatenar os mapas de características em diferentes escalas, conforme ilustrado nas Figuras 37 e 40. O *pooling* imita o sistema visual humano, em que são executadas operações de redução de dimensionalidade (redução da resolução) para representar características da imagem em um nível mais alto de abstração, tornando o mapa de características menor e simplificando a complexidade computacional da rede, uma vez que são compactadas e extraídas as principais características da imagem (ZHU et al., 2021).

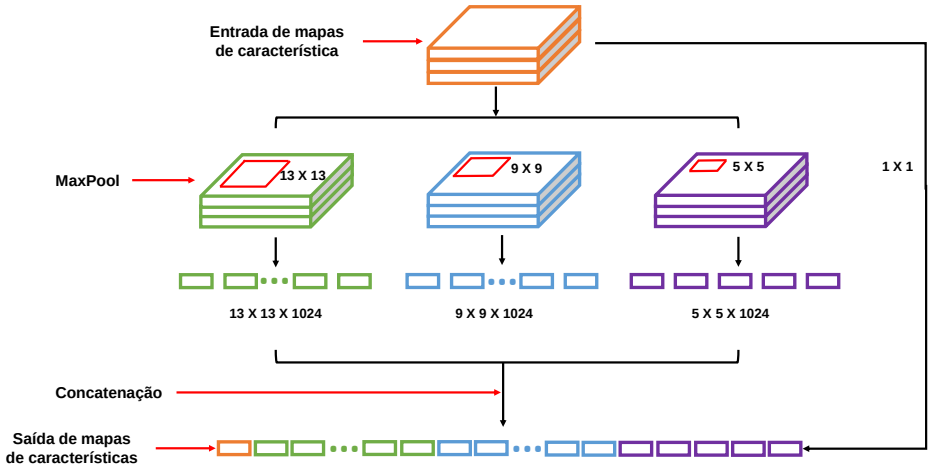


Figura 40 – Estrutura do módulo SPP do *Backbone* da rede. Foi utilizado o método de agrupamento *MaxPool* com agrupamentos de tamanho igual a 1×1, 5×5, 9×9 e 13×13, seguido da operação *Concat* para concatenar os mapas de características em diferentes escalas.

A estrutura do *Backbone* é responsável por extrair os mapas de características de diferentes tamanhos da imagem de entrada por meio de múltiplas convoluções e agrupamentos (ZEILER; TAYLOR; FERGUS, 2011). A estrutura do *Neck*, por sua vez, é responsável pela fusão destes mapas de características obtidos de diferentes níveis da arquitetura para obter mais informações contextuais e reduzir problemas relacionados à perda de informações durante o processo de extração de características das imagens (ZHU et al., 2021).

No processo de fusão dos mapas de características provenientes do *Backbone*, a *Feature Pyramid Network* (FPN) (LI et al., 2019) e o PAN (*Path Aggregation Network*) (LIU et al., 2018) são utilizados conforme ilustrado na Figura 41.

Segundo Li et al. (2019), FPN é um componente básico em sistemas de reconhecimento para detectar objetos em diferentes escalas, pois trata-se de uma arquitetura *top-down* com conexões laterais que constrói mapas de características com significado semântico de alto nível em todas as escalas. Entretanto, a estrutura do FPN

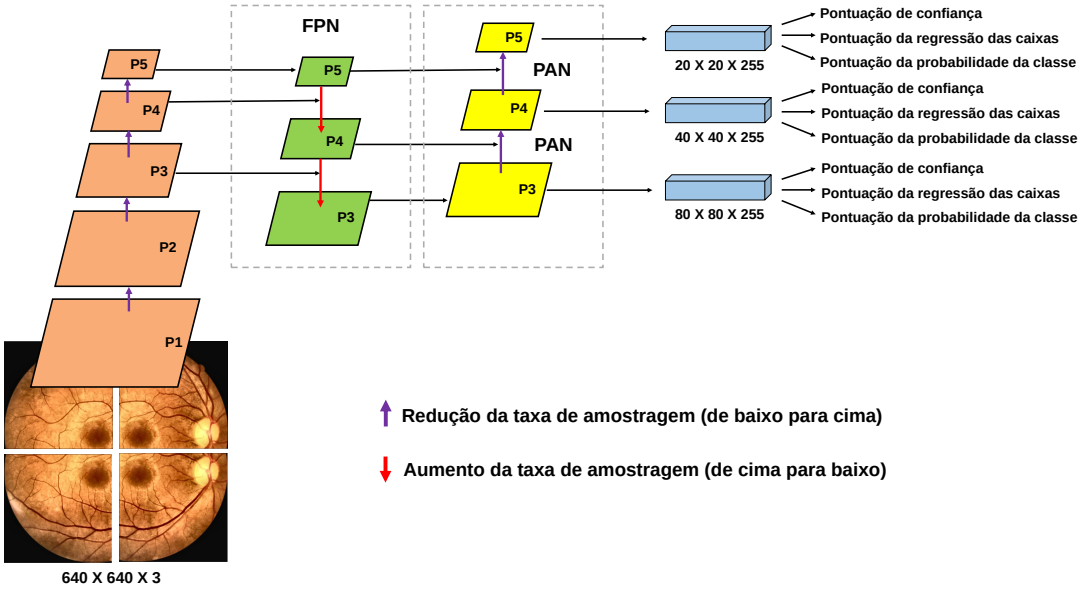


Figura 41 – Estrutura FPN+PAN utilizada no *Neck* da arquitetura da rede neural da abordagem. As duas estruturas utilizadas em conjunto reforçam a capacidade de fusão de características da estrutura do *Neck*. A detecção das lesões é realizada nas camadas P3, P4 e P5 da estrutura FPN+PAN, possuindo saídas com tamanho de 80×80×255, 40×40×255 e 20×20×255, respectivamente.

por si só pode ser excessivamente extensa, pois a informação espacial pode precisar ser propagada para centenas de camadas. Nesse contexto, utilizou-se uma FPN em conjunto com uma PAN. A estrutura do PAN segue um caminho ascendente adicional em relação ao caminho descendente percorrido pelo FPN, auxiliando a encurtar esse caminho usando conexões laterais como uma conexão de atalho.

A arquitetura PAN transmite características de localização fortes dos mapas de características inferiores para os mapas de características superiores. As duas estruturas associadas reforçam a capacidade de fusão de características da estrutura do *Neck* (ZHU et al., 2021).

Na estrutura do *Neck* foram utilizados 4 módulos CSP (C3), conforme ilustrado na Figura 35. Estes módulos C3 foram adotados para fortalecer a capacidade de integração das características extraídas das lesões durante a propagação destas informações na rede neural. Na Figura 41 é possível observar que a detecção das lesões é realizada nas camadas P3 (objetos pequenos), P4 (objetos médios) e P5 (objetos grandes), com tamanhos de 80×80×255, 40×40×255 e 20×20×255, respectivamente.

Por fim, o *Head* é a parte da rede neural encarregada por fazer a previsão densa (previsão final), sendo composta por um vetor que contém a caixa delimitadora predita (coordenadas centrais, altura, largura), a pontuação de confiança da previsão e o rótulo da classe a qual o objeto detectado pertence. O mecanismo de predição usado no *Head* da arquitetura da rede neural profunda da abordagem é equivalente ao utilizado

no YOLOv3 (REDMON; FARHADI, 2018). YOLO é um modelo de detecção de objetos de estágio único, pois não utiliza uma etapa específica para a classificação de regiões de interesse na imagem, sendo a classificação e a detecção realizadas em estágio único (REDMON et al., 2016; ZHENG; ZHAO; LI, 2021).

YOLO é uma arquitetura de detecção de objetos de ponta a ponta que oferece vantagens na otimização de tarefas de detecção e velocidade de inferência em comparação com modelos de dois estágios que primeiro fazem a previsão de regiões de interesse para depois a classificação dos objetos. O modelo YOLO trata a detecção de objetos como um problema de regressão e as classes como probabilidades condicionais (NGUYEN et al., 2020; MAMDOUH; KHATTAB, 2021; IYER et al., 2021). Para tanto, o YOLO divide a imagem em células de grade S$\times$S. Cada célula da grade prevê um conjunto de caixas delimitadoras (*bounding box regression*), a confiança de um objeto existir na célula da grade (*objectness*) e a classe do objeto (*class probability*) (REDMON et al., 2016). A saída do YOLO é um tensor na forma ($S \times S \times N^o$ de filtros) (MAMDOUH; KHATTAB, 2021), de modo que o número de filtros é dado pela Equação 25:

$$filters = (classes + P_c + X_{center} + Y_{center} + width + height) * B \tag{25}$$

onde: $classes$ é o número de classes de treinamento; P_c é a confiança de um objeto existir nesta célula da grade; X_{center} e Y_{center} são as coordenadas x e y da caixa delimitadora em relação à célula da grade, respectivamente; *width* e *height* são a largura e altura da caixa delimitadora em relação à célula da grade; e, B é o número de caixas preditas para cada célula da grade.

O conjunto de caixas delimitadoras preditas por cada célula da grade são conhecidas como caixas de ancoragem. A arquitetura prevê quatro coordenadas de deslocamento (t_x, t_y, t_w, t_h) para cada caixa delimitadora. A largura e altura do valor anterior da caixa delimitadora são correlacionados linearmente. (c_x, c_y) são as coordenadas de deslocamento da parte superior do canto esquerdo da imagem e (p_w, p_h) são a largura e altura da caixa delimitadora (ZHENG; ZHAO; LI, 2021; MAMDOUH; KHATTAB, 2021). As coordenadas preditas (b_x, b_y, b_w, b_h) são obtidas conforme as Equações:

$$b_x = (2 * \sigma(t_x) - 0.5) + c_x \tag{26}$$

$$b_y = (2 * \sigma(t_y) - 0.5) + c_y \tag{27}$$

$$b_w = p_w * (2 * \sigma(t_w))^2 \tag{28}$$

$$b_h = p_h * (2 * \sigma(t_h))^2 \tag{29}$$

Cada célula da grade prevê a confiança de um objeto existente nela. Essa pontuação de confiança reflete a probabilidade de a caixa conter um objeto (*objectness*). Além disso, também é calculada a pontuação de confiança da classe como (ZHENG; ZHAO; LI, 2021; MAMDOUH; KHATTAB, 2021):

$$confidence = P(object) * IoU \tag{30}$$

$$class\_confidence = P(classe|object) * confidence \tag{31}$$

onde: IoU (*Intersection Over Union*) é a razão entre a área de interseção e a área de união das caixas delimitadoras preditas e a caixa delimitadora do *Ground Truth*.

Cada objeto é previsto por uma caixa delimitadora e no caso de serem detectadas várias caixas delimitadores para um mesmo objeto então aplicou-se a técnica de NMS, que permite descartar as caixas delimitadoras que tenham um IoU abaixo de um limiar pré-definido, conforme apresentado na Tabela 11.

A estrutura do *Head* utilizada na abordagem é composta por 3 camadas responsáveis pela realização da detecção das lesões de fundo, sendo que cada uma destas camadas divide a imagem em células de grade de tamanhos 20×20×255, 40×40×255 e 80×80×255, conforme ilustrado nas Figuras 35 e 41. Os tamanhos 20×20, 40×40 e 80×80 são obtidos após as reduções de resolução realizadas nos módulos *Focus* e *Conv* do *Backbone* da arquitetura e a quantidade de filtros (canais) igual a 255 é obtida por meio da Equação 32:

$$filters = (classes + 5) * B \tag{32}$$

Quanto menor o tamanho dos mapas de características, maior será a área da imagem à qual cada unidade da grade no mapa de características corresponde, indicando que é adequado para detectar objetos grandes a partir dos mapas de características de tamanho 20×20×255, ao passo que mapas de características de tamanho 80×80×255 são mais adequados para detectar objetos pequenos.

A Tabela 10 apresenta os parâmetros utilizados nos diversos módulos que compõem a arquitetura da rede neural profunda da abordagem. É possível observar que é realizada uma redução de amostragem das imagens de entrada nos módulos 1 (*Focus*), 2 (Conv), 4 (Conv), 6 (Conv) e 8 (Conv) (vide Figura 35), por exemplo, ficando com os tamanhos 320×320, 160×160, 80×80, 40×40 e 20×20 *pixels*, respectivamente. Também é possível constatar que a detecção das lesões é realizada nos módulos 25 (Conv2d), 26 (Conv2d) e 27 (Conv2d) localizados no *Head* da arquitetura,

e que as dimensões dos mapas de características são 80× 80×255, 40×40×255 e 20×20×255, respectivamente.

Tabela 10 – Parâmetros utilizados nos diversos módulos que compõem a arquitetura da rede neural profunda da abordagem.

Backbone					
Id	**Módulo**	**Tamanho da imagem**	**Número de *kernels* convolucionais**	**Tamanho do *kernel***	***Stride***
1	*Focus*	320×320	64	1×1	1
2	Conv	160×160	128	3×3	2
3	CSP (C3)	160×160	128	-	-
4	Conv	80×80	256	3×3	2
5	CSP (C3)	80×80	256	-	-
6	Conv	40×40	512	3×3	2
7	CSP (C3)	40×40	512	-	-
8	Conv	20×20	1024	3×3	2
9	SPP	20×20	1024	-	-
10	CSP(C3)	20×20	1024	-	-
Neck					
11	Conv	20×20	512	1×1	1
12	*Upsample*	40×40	512	-	-
13	Concat	40×40	1024	-	-
14	CSP (C3)	40×40	512	-	-
15	Conv	40×40	256	1×1	1
16	*Upsample*	80×80	256	-	-
17	Concat	80×80	512	-	-
18	CSP (C3)	80×80	256	-	-
19	Conv	40×40	256	3×3	2
20	Concat	40×40	512	-	-
21	CSP (C3)	40×40	512	-	-
22	Conv	20×20	512	3×3	2
23	Concat	20×20	1024	-	-
24	CSP (C3)	20×20	1024	-	-
Head					
25	Conv2d (P3)	80×80	256	1×1	1
26	Conv2d (P4)	40×40	512	1×1	1
27	Conv2d (P5)	20×20	1024	1×1	1

Segundo Taqi et al. (2018), uma rede neural precisa que as variáveis de cada camada sejam alteradas de forma que esta funcione melhor no processo de classificação. Para esta tarefa é importante que o desempenho da rede seja continuamente medido através da comparação entre o resultado obtido e o resultado esperado. Portanto, o objetivo da otimização é minimizar o resultado da função `cross-entropy()`, que é sempre positivo e se torna zero quando o resultado obtido é exatamente o mesmo resultado esperado.

Ao realizar o treinamento de uma rede neural um dos maiores desafios é a escolha correta dos hiperparâmetros. Em geral, é necessária a realização de inúmeros testes a fim de obter o ajuste mais adequado de hiperparâmetros para que a rede neural profunda alcance melhores resultados. Neste contexto, a escolha do método

de otimização tem um impacto significativo no comportamento da rede neural (RIBEIRO; JUNIOR, 2020). Para que os otimizadores obtenham um bom desempenho, os hiperparâmetros devem ser ajustados adequadamente. Por exemplo, pequenas mudanças na taxa de aprendizado podem alterar drasticamente o desempenho da rede neural (RIBEIRO; JUNIOR, 2020; CHEN et al., 2021).

Um componente chave para o ajuste de hiperparâmetros é a seleção de uma boa taxa de aprendizado, e possivelmente o *Momentum* (RIBEIRO; JUNIOR, 2020; CHEN et al., 2021). Em um esforço para diminuir o custo de treinamento de *Deep Neural Networks* (DNNs), métodos adaptativos baseados em gradiente (DUCHI; HAZAN; SINGER, 2011; KINGMA; BA, 2017; MA; YARATS, 2019) foram desenvolvidos. Porém, existem casos onde o uso destes métodos levam a um degradação de desempenho da rede (SHAH; KYRILLIDIS; SANGHAVI, 2020; WILSON et al., 2018), mas isso pode ser resultado de um ajuste deficiente dos hiperparâmetros (AGARWAL et al., 2020; SHAH; KYRILLIDIS; SANGHAVI, 2020; S et al., 2020; CHOI et al., 2020).

De acordo com Ruder (2017), o algoritmo de Gradiente Descendente é um dos métodos mais populares utilizados na tarefa de otimização de redes neurais. Nesse método é possível minimizar uma função objetivo $J(\theta)$ pelos parâmetros $\theta \in \mathbb{R}^d$ de um modelo através da atualização desses parâmetros na direção oposta do gradiente da função objetivo $\nabla\theta\ J(\theta)$ com relação aos parâmetros. A taxa de aprendizado η determina o tamanho dos passos para atingir o mínimo local. Como função de otimização realizaram-se experimentos com os otimizadores SGD (*Stochastic Gradient Descent*) (RUDER, 2016) com *Momentum* e Adam (KINGMA; BA, 2017).

O método SGD com *Momentum* e Adam são os otimizadores mais utilizados atualmente para o treinamento de DNNs em função de sua estabilidade e queda menos drástica de desempenho ao longo do treinamento da rede neural (RIBEIRO; JUNIOR, 2020; CHEN et al., 2021). O SGD, proposto por Sutskever et al. (2013), executa uma atualização de parâmetros para cada amostra de treino $x^{(1)}$ e cada rótulo $\theta = \theta - \eta * \nabla_\theta J(\theta; x^{(1)}, y^{(1)})$. Como o SGD executa uma atualização de parâmetro por vez, de modo geral, seu algoritmo é muito eficiente, podendo inclusive ser utilizado para aprendizado *online*. Esse algoritmo realiza atualizações com alta variância que leva a função objetivo sofrer grandes flutuações (RUDER, 2017). O *Momentum*, por sua vez, é um método que ajuda a acelerar o otimizador SGD a encontrar a direção correta de convergência mais rapidamente, além de minimizar oscilações do otimizador. Isso ocorre através da adição de uma fração γ do vetor de atualização do passo anterior ao atual vetor de atualização (RUDER, 2017).

Proposto por Kingma; Ba (2017), o Adam é um método de otimização que calcula taxas de aprendizado adaptativas para cada parâmetro. Segundo Taqi et al. (2018), o algoritmo atualiza as médias móveis exponenciais do gradiente m_t (Equação 33) e do gradiente quadrado v_t (Equação 34), onde m_t e v_t são, respectivamente, as esti-

mações do primeiro momento e segundo momento dos gradientes e β_1 e β_2 são os hiperparâmetros que controlam as taxas de decaimento exponencial dessas médias móveis pertencentes ao intervalo [0,1].

$$m_t = \beta_1 m_{t-1} + (1 - \beta_1) g_t \tag{33}$$

$$v_t = \beta_2 v_{t-1} + (1 - \beta_2) g_t^2 \tag{34}$$

Segundo Ribeiro; Junior (2020), a regra de atualização dos parâmetros é dada pela Equação 35 (RIBEIRO; JUNIOR, 2020):

$$\theta_{t+1} = \theta_t - \frac{\eta}{\sqrt{\hat{v}_t + \epsilon}} \hat{m}_t \tag{35}$$

A função de perda final utilizada na abordagem é calculada com base na pontuação de confiança (*Objectness*), na pontuação de classificação (*Class Probability*) e na pontuação de regressão das caixas delimitadoras (*Bounding Box Regression*), conforme a Equação 36. A *Objectness* determina se há objetos na célula da grade, *Class Probability* determina a qual categoria os objetos que estão em uma célula de grade pertencem e *Bounding Box Regression* é calculada apenas quando a caixa predita contém objetos. Neste caso, o cálculo da *Bounding Box Regression* é realizado mediante a comparação da caixa predita com a caixa associada ao *Ground Truth* do objeto detectado.

$$Loss = L_{Objectness} + L_{ClassProbability} + L_{BoundingBoxRegression} \tag{36}$$

Para calcular as funções de perda de pontuação de confiança (*objectness*) e de pontuação de classificação (*class probability*) utilizamos *Binary Cross-Entropy* com a função Logits do PyTorch (JADON, 2020). *Binary Cross-Entropy* descreve a distância entre duas distribuições de probabilidade, ou seja, quanto menor a entropia cruzada, mais próximas estão as duas distribuições de probabilidade. Assim como no YOLOv3 (REDMON; FARHADI, 2018) foi utilizada a entropia cruzada binária para reduzir a complexidade computacional do modelo. Para calcular a função de perda referente à regressão de caixa delimitadora utilizou-se a função de perda *Generalized Intersection over Union* ($GIoU$) (REZATOFIGHI et al., 2019; LIN et al., 2019; OKSUZ et al., 2021).

A perda de IoU ($LIoU$) é uma função de regressão das caixas delimitadoras invariável à escala (ZHANG et al., 2019), conforme a Equação 37, em que B_p é a caixa delimitadora predita e B_{gt} é caixa delimitadora com o *Ground Truth* do objeto.

$$L_{IoU} = 1 - \frac{|B_p \cap B_{gt}|}{|B_p \cup B_{gt}|} \tag{37}$$

A principal desvantagem da $LIoU$ é que a IoU assume o valor zero quando não há sobreposição entre as caixas delimitadoras, gerando falha ao indicar a que distância estas caixas estão separadas umas das outras (REZATOFIGHI et al., 2019), podendo causar dissipação de gradiente ao treinar um modelo, por exemplo. Para contornar este problema, Rezatofighi et al. (2019) apresentaram uma função de perda $GIoU$, conforme a Equação 38:

$$L_{GIoU} = 1 - IoU + \frac{|C - B_p \cup B_{gt}|}{|C|} \tag{38}$$

$GIoU$ é uma função de perda que tem um termo de penalidade C junto com a função de perda de IoU $(1 - IoU)$, onde C é a menor caixa cobrindo B_p e B_{gt}. Devido à introdução do termo de penalidade C, a perda de $GIoU$ continua expandindo o tamanho da caixa predita até que se sobreponha à caixa de destino e, então, o termo IoU funcionará para maximizar a área de sobreposição da caixa delimitadora (REZATOFIGHI et al., 2019).

Para duas caixas B_p e B_{gt}, encontra-se a menor caixa (C) que inclui B_p e B_{gt}. Em seguida, é calculada a razão entre o volume ocupado por C excluindo $B_p \cup B_{gt}$ e dividindo-o pela área total ocupada por C. Dessa forma, o modelo pode trabalhar na distância vazia entre duas caixas e, assim, aproxima a caixa predita da caixa com o *Ground Truth*, reduzindo o espaço vazio. De forma análoga à função de perda IoU, a $GIoU$ também é invariante à escala (REZATOFIGHI et al., 2019; ZHENG et al., 2020).

No pós-processamento da detecção das lesões de fundo foi necessário realizar a triagem e remoção de caixas delimitadoras duplicadas que representam o mesmo objeto. Para tanto, foi utilizada a técnica de NMS, que mantém a caixa delimitadora detectada com maior índice de precisão. Utilizou-se o método de NMS baseado no valores de IoU obtidos (`IoU_nms`), em que foi configurado um limiar de 0,45 (ZHU et al., 2021) para a etapa de treinamento.

Dentre as contribuições que se propôs realizar na detecção das lesões de fundo, destacam-se a aplicação de técnicas de pré-processamento nas imagens para minimizar a presença de *outliers* e realçar características das lesões, assim como o *cropping* parcial do plano de fundo preto das imagens de fundo (retina) para minimizar a geração de falsos positivos. Além disso, dividiu-se as imagens em blocos (*tiles*) para serem utilizados no treinamento, validação e teste da abordagem, com o objetivo de passar à camada de entrada da rede neural imagens com a maior resolução possível, a fim de minimizar a perda de informações ocasionadas pelo redimensionamento das imagens originais. A aplicação de *Tilling* possibilitou aumentar o campo receptivo das imagens utilizadas para treinamento da rede neural, minimizando a perda de detalhes fundamentais para a detecção das microlesões como os microaneurismas.

Realizou-se o aumento de dados por meio de técnicas de última geração (*Mosaic*, *MixUp*, *Copy-Paste* e Transformações Geométricas) a fim de minimizar o subajuste relacionado à quantidade limitada de lesões de fundo rotuladas. Também realizou-se a divisão do conjunto de dados em treinamento, validação e teste para ajuste e validação da abordagem em diferentes conjuntos de imagens de fundo, a fim de evitar o ajuste excessivo do modelo aos dados e realizar uma avaliação efetiva da capacidade preditiva da abordagem sobre dados não conhecidos *a priori*.

O cálculo para otimização das âncoras utilizadas para a detecção das lesões foi realizado de forma automática, por meio de um algoritmo de aprendizado não supervisionado K-means e um Algoritmo Genético. A otimização das âncoras proporcionou uma melhoria na localização das caixas delimitadoras das lesões de fundo, pois ao invés da detecção ser realizada com base nos tamanhos de âncoras calculadas no conjunto de dados COCO, o ajuste dos tamanhos das caixas de âncora foi realizado com base nas anotações das lesões de fundo do conjunto de dados utilizado para treinamento da rede neural profunda utilizada na abordagem.

Por fim, as contribuições relacionadas à arquitetura da rede neural profunda da abordagem incluem: utilização da função de ativação SiLU em toda a rede neural para tornar a arquitetura mais eficiente; uso da função de perda *Bounding Box Regression* $GIoU$ para viabilizar o cálculo de perdas para caixas delimitadoras não sobrepostas; utilização de módulos CSP (C3) no *Head* e *Neck* para reforçar a propagação e reutilização de características das lesões ao longo da rede neural a fim de minimizar problemas de dissipação de gradiente e reduzir a quantidade de parâmetros da arquitetura; e, utilização de FPN e PAN de forma integrada no *Neck* da arquitetura para melhorar a capacidade de propagação das características extraídas das lesões e possibilitar detecções em diferentes escalas.

6.2.1 Pré-treinamento

Transferência de aprendizado é um método empregado na área de aprendizado de máquina que consiste em reutilizar informações aprendidas em uma determinada tarefa como ponto de partida para a solução de uma nova tarefa (PAN; YANG, 2010). Esse método costuma ser utilizado quando não é possível obter um conjunto de dados de larga escala com objetos rotulados para resolver uma determinada tarefa de Visão Computacional (BLITZER; DREDZE; PEREIRA, 2007). Portanto, é importante que alguns fatores sejam considerados para a adoção deste método, tais como: (1) o tamanho do conjunto de dados que será utilizado na tarefa-alvo; e, (2) a similaridade entre o conjunto de dados utilizado para treinamento dos pesos que serão reutilizados e o conjunto de dados da tarefa-alvo.

Caso durante o treinamento da tarefa-alvo nenhum peso pré-treinado é mantido fixo e a arquitetura da rede neural usada para resolver a tarefa-alvo é a mesma utilizada para gerar os pesos pré-treinados, dá-se o nome de *Transfer Learning*. Caso os pesos de um treinamento realizado anteriormente sejam mantidos fixos (congelados para determinadas camadas) e o treinamento seja realizado novamente para ajustar as camadas alteradas, dá-se o nome de *Fine-Tuning* (PAN; YANG, 2010). O ajuste fino de redes pré-treinadas é um método de *Transfer Learning* (MAMDOUH; KHATTAB, 2021). A utilização de pesos pré-treinados baseia-se no reuso de padrões aprendidos principalmente nas primeiras camadas de uma rede neural, tais como: linhas, contornos, cores, etc. Assim, estes pesos das camadas iniciais contém informações (características) que podem ser reutilizadas em diferentes tipos de tarefas que envolvam a classificação de objetos em imagens.

A abordagem é composta por um detector *Single-Stage* baseado em algoritmos de aprendizado profundo de última geração. Embora estes algoritmos apresentem alta precisão, necessitam de conjuntos de dados com grande quantidade de exemplos (MAMDOUH; KHATTAB, 2021). Nesse contexto, como os conjuntos de dados públicos de RD não contém grande quantidade de lesões rotuladas, além de uma etapa de aumento de dados, a abordagem conta com uma etapa de pré-treinamento. Nesta etapa, aplica-se *Transfer Learning* com os pesos pré-treinados no conjunto de dados COCO (LIN et al., 2014) e *Fine-Tuning* das últimas camadas da rede neural com o conjunto de dados de fundo utilizado nos experimentos. Para realizar o ajuste fino da abordagem mantiveram-se os pesos das primeiras camadas e alteraram-se apenas os pesos das últimas camadas da rede neural.

COCO fornece um grande conjunto de dados com imagens rotuladas para tarefas de detecção de objetos. Modificou-se a saída da rede neural da abordagem para se adequar ao problema de detecção das lesões de fundo, preservando o conhecimento (pesos) das camadas iniciais. A reutilização de informações destas camadas iniciais é fundamental para a obtenção de características mais básicas das lesões de fundo, tais como: contornos, arestas, etc. Além disso, o pré-treinamento possibilitou uma redução do custo computacional e tempo de treinamento da abordagem. O método adotado para realizar a transferência de aprendizagem baseou-se no trabalho proposto por Franke et al. (2021) e consiste das quatro etapas apresentadas a seguir:

1. As camadas iniciais da arquitetura da abordagem, focadas na detecção de características mais básicas de objetos, foram pré-treinadas com os pesos do conjunto de dados COCO, composto por 80 classes.

2. As últimas três camadas (de um total de 283 camadas) que compõem o *Head* da arquitetura da abordagem são cortadas e substituídas por novas camadas.

3. As novas camadas adicionadas são ajustadas por meio do treinamento da rede

neural no conjunto de dados de RD, enquanto os pesos das camadas iniciais são congelados.

4. Após o ajuste fino das camadas do *Head* da arquitetura, toda a rede neural é descongelada e treinada novamente, de forma que pequenos ajustes nos pesos sejam realizados em toda a rede.

O ajuste fino da rede neural teve como objetivo realizar a otimização da abordagem a fim de alcançar resultados mais precisos. Para tanto, fizeram-se também ajustes em hiperparâmetros como tamanho do lote, número de épocas, taxa de aprendizado, etc. Segundo Franke et al. (2021), a otimização dos hiperparâmetros visa encontrar um conjunto de valores que produza um modelo ideal, em que uma função de perda pré-definida é minimizada. Assim como no trabalho proposto por Franke et al. (2021), a metodologia adotada para realizar o ajuste fino de hiperparâmetros da abordagem consistiu das seguintes etapas:

1. Para cada ajuste realizado, um valor de hiperparâmetro é variado e a abordagem é retreinada, mantendo constantes os demais valores de hiperparâmetros.

2. O efeito desta mudança é analisado por meio da avaliação de desempenho da abordagem com as métricas *Average Precision* (AP) e *mean Average Precision* (mAP), que serão apresentadas e discutidas na próxima Seção.

3. Se houver uma melhoria nos valores das métricas, o valor do hiperparâmetro é ainda ajustado (aumentado ou diminuído) até que o máximo local seja alcançado.

4. O mesmo processo é realizado para os demais hiperparâmetros até que seja obtido um conjunto de valores que produzam os resultados de AP e mAP que são apresentados nesta tese. Para realizar o cálculo do AP foi utilizada a mesma abordagem empregada no desafio COCO (COCO, 2021).

Após a realização destas etapas utilizando o conjunto de dados de validação, obtidos do DDR, foi encontrado o ajuste ideal de hiperparâmetros, conforme apresenta-se na Tabela 11. Com os valores de hiperparâmetros devidamente ajustados, a próxima etapa foi avaliar a abordagem no conjunto de dados de teste do conjunto de dados utilizado nos experimentos.

6.3 Experimentos, Resultados e Discussões

Para avaliar o desempenho da abordagem, realizaram-se experimentos utilizando o conjunto de dados público de Retinopatia Diabética DDR. Para evitar enviesamento dos resultados dividimos o conjunto de dados em conjunto de treino, validação e teste,

Tabela 11 – Hiperparâmetros ajustados durante a etapa de validação utilizando o conjunto de dados DDR.

Hiperparâmetro	**Valor**
Dropout	10%
Early Stopping	*Patience value* $=$ 100
Função de Ativação	SiLU
Limiar IoU NMS	0,45
Limite de Confiança	0,25
Momentum	0,937
Número de Épocas	8.000
Otimizador	SGD e Adam
Tamanho das âncoras iniciais (COCO)	(10, 13), (16, 30), (33, 23) - P3 (30, 61), (62, 45), (59, 119) - P4 (116, 90), (156, 198), (373, 326) - P5
Tamanho das âncoras ajustadas	(3, 3), (4, 4), (7, 7) - P3 (10, 10), (15, 15), (23, 28) - P4 (33, 24), (44, 49), (185, 124) - P5
Tamanho do Lote	32
Taxa de Aprendizado	0,01
Weight Decay	0,0005

em uma proporção de 50%, 20% e 30%, respectivamente. A arquitetura foi implementada e treinada com base no modelo YOLOv5. Primeiramente, realizou-se transferência de aprendizado com base nos pesos pré-treinados no conjunto de dados COCO, em seguida, o ajuste fino nas camadas de detecção da arquitetura e, por fim, retreinou-se toda a rede neural.

Durante o treinamento da abordagem utilizou-se o método de regularização *Early Stopping* (PRECHELT, 1998). Com o uso desta técnica não foi necessário definir estaticamente a quantidade de épocas necessárias para treinamento da abordagem, uma vez que no final de cada época é calculada a precisão da classificação nos dados de validação e quando a precisão para de melhorar, o treinamento é terminado. Com o uso de *Early Stopping* foi possível evitar problemas como: *underfitting*, em que a rede neural não consegue extrair características suficientes das imagens durante o treinamento em função de uma quantidade insuficiente de épocas; e, *overfitting*, em que a rede neural se ajusta excessivamente aos dados de treinamento em função de uma quantidade excessiva de épocas (ZHANG et al., 2017). Para tanto, definiu-se um parâmetro para terminar o treinamento se a precisão da classificação não melhorasse durante as últimas 100 épocas, conforme apresentado na Tabela 11.

Além do método *Early Stopping* utilizou-se a técnica *Dropout* (SRIVASTAVA et al., 2014) para auxiliar na regularização da abordagem. Esta técnica é amplamente utilizada para regularizar o treinamento de redes neurais profundas (LIANG et al., 2021). *Dropout* auxilia na regularização do modelo (LABACH; SALEHINEJAD; VALAEE, 2019), sem modificar a função de custo. Ainda, com o uso de *Dropout* alguns neurônios ocultos da rede neural são desligados aleatoriamente e de forma temporária,

sem a alteração dos neurônios de entrada e saída. Portanto, esta técnica faz com que alguns neurônios não funcionem de acordo com uma determinada probabilidade durante o treinamento.

Dropout auxilia na regularização porque reduz co-adaptações complexas de neurônios, fazendo que alguns neurônios sejam forçados a aprender características que *a priori* seriam aprendidas por outros neurônios da arquitetura. Em suma, a ideia principal é descartar unidades (neurônios) aleatoriamente (junto com suas conexões) da rede neural durante o treinamento, evitando que as unidades se adaptem demasiadamente aos dados (SRIVASTAVA et al., 2014), reduzindo a possibilidade de problemas relacionados a *overfitting* da rede neural após o aumento de dados, por exemplo. O parâmetro definido na abordagem para utilização de *Dropout* foi apresentado na Tabela 11.

Avaliou-se a abordagem com as métricas AP e mAP para comparar os resultados. Estas métricas são frequentemente utilizadas para medir a precisão de algoritmos de aprendizado profundo que realizam a detecção de objetos (KONISHI et al., 2016; REDMON; FARHADI, 2018). Comparou-se a abordagem com abordagens correlatas encontradas na literatura, incluindo SSD (LI et al., 2019), YOLO (LI et al., 2019), YOLOv3 (ALYOUBI; ABULKHAIR; SHALASH, 2021) e YOLOv4 (SANTOS et al., 2021), conforme apresentado na Tabela 12. Após os experimentos realizados durante a etapa de validação da abordagem no conjunto de dados DDR utilizando o otimizador SGD, o melhor resultado foi obtido utilizando o método de *Tilling*, com um mAP de 0,2490 (indicado em negrito) e valores de AP com limite de IoU de 0,5 iguais a 0,2290, 0,3280, 0,1050 e 0,3330, para Exsudatos Duros (EX), Exsudatos Algodonosos (SE), Microaneurismas (MA) e Hemorragias (HE), respectivamente, conforme apresentado na Tabela 12.

Tabela 12 – Resultados obtidos pela abordagem com otimizador SGD em comparação aos trabalhos relacionados com as métricas AP e mAP para o limite de IoU de 0,5 no conjunto de validação do conjunto de dados DDR.

Modelos	AP				mAP
	EX	**SE**	**MA**	**HE**	
SSD (LI et al., 2019)	0	0,0227	0	0,0007	0,0059
YOLO (LI et al., 2019)	0,0039	0	0	0,0101	0,0035
YOLOv3+SGD (ALYOUBI; ABULKHAIR; SHALASH, 2021)	-	-	-	-	0,1100
YOLOv3+SGD+*Dropout* (ALYOUBI; ABULKHAIR; SHALASH, 2021)	-	-	-	-	0,1710
YOLOv4 (SANTOS et al., 2021)	0,0370	0,1493	0,0193	0,0849	0,0716
YOLOv5-S (*baseline*) (JOCHER, 2020)	0,1360	0,2600	0,0584	0,2200	0,1686
Abordagem para Detecção+SGD sem *Tilling*	0,1490	0,4060	0,0454	0,2780	0,2200
Abordagem para Detecção+SGD com *Tilling*	0,2290	0,3280	0,1050	0,3330	**0,2490**

Para investigar os resultados da abordagem, optou-se por analisar a curva PR, em detrimento da curva ROC (*Receiver Operating Characteristic*) (DAVIS; GOADRICH, 2006), uma vez que a curva ROC não é recomendada para situações em que o conjunto de dados apresenta desbalanceamento da quantidade de exemplos entre as classes investigadas. Nestes casos, a curva ROC costuma apresentar um AUC muito elevado em função do preditor classificar corretamente a classe com maior número de exemplos (classe majoritária) (MANNING; RAGHAVAN; SCHÜTZE, 2008; FLACH; KULL, 2015).

Para se avaliar os resultados obtidos utilizou-se Precisão e Revocação, que são métricas de desempenho normalmente utilizadas para avaliar sistemas de classificação de imagens e recuperação de informação. De modo geral, a precisão e a revocação não são discutidas isoladamente e há problemas que podem requerer uma Revocação mais alta em relação à Precisão, ou vice-versa, dependendo da importância dada aos falsos negativos versus falsos positivos. Em problemas de classificação envolvendo imagens médicas, por exemplo, geralmente o que se deseja é minimizar a incidência de falsos negativos, portanto, uma Revocação elevada torna-se mais importante que uma Precisão elevada, uma vez que um falso negativo pode implicar em um diagnóstico médico errado e, portanto, riscos à saúde do paciente.

A Figura 42 apresenta o gráfico referente à curva PR com limite de IoU de 0,5 obtido durante a etapa de validação utilizando a abordagem com otimizador SGD e *Tilling* no conjunto de dados DDR. No gráfico são plotados os valores de AP obtidos pelas lesões de fundo, conforme os resultados apresentados na Tabela 12, cujo valor de *mean Average Precision* obtido por todas as classes corresponde a 0,249.

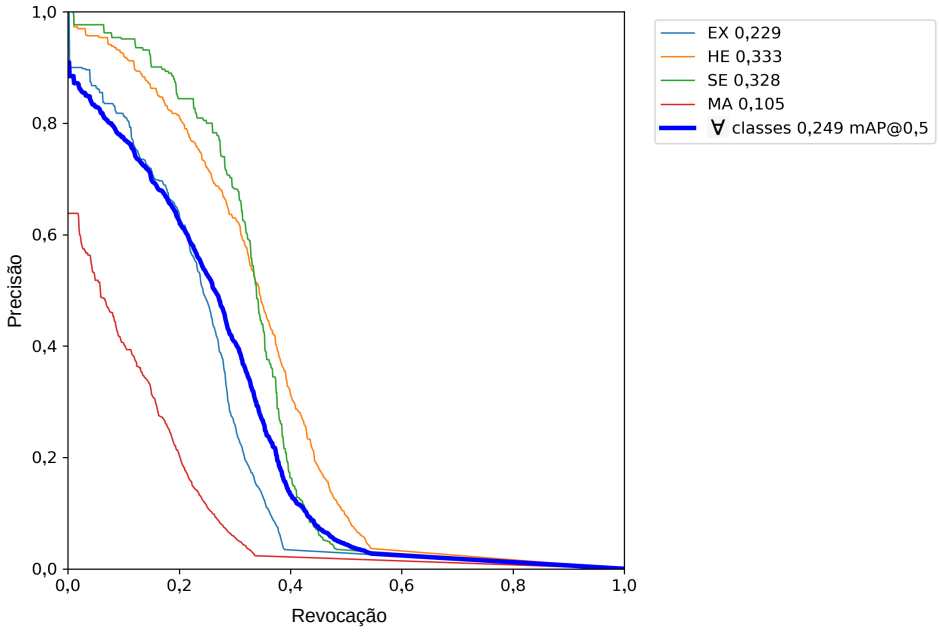


Figura 42 – Gráfico com a curva PR com limite de IoU de 0,5 obtido durante a etapa de validação da abordagem com otimizador SGD e *Tilling* no conjunto de dados DDR.

O eixo x da curva PR representa a Revocação ao passo que o eixo y representa a Precisão. Esta curva foca principalmente no desempenho das classes positivas, que é fundamental quando se lida com classes desequilibradas. Desse modo, no espaço PR, o objetivo é estar no canto superior direito (1, 1), significando que o preditor classificou todos os positivos como positivos (Revocação $= 1$), e que tudo que foi classificado como positivo é verdadeiro positivo (Precisão $= 1$).

Com base na análise do gráfico da curva PR é possível verificar que a abordagem encontrou maior dificuldade em predizer Microaneurismas (MA) (curva na cor vermelha) seguido dos Exsudatos Duros (EX) (curva na cor ciano), sendo os melhores resultados obtidos com a predição de Exsudatos Algodonosos (SE) (curva na cor verde) e Hemorragias (HE) (curva na cor laranja), respectivamente. A baixa precisão obtida na detecção dos MA está relacionada principalmente com o tamanho destas microlesões e à dissipação de gradiente destes objetos quando a rede neural é treinada, ocasionando uma taxa elevada de erros, conforme pode ser observado na matriz de confusão apresentada na Figura 43, com 79% de *background* FN e 38% de *background* FP, ficando atrás neste quesito apenas dos exsudatos duros, com 40%. O fato da abordagem ter alcançado melhores resultados na predição dos SE está associado às características morfológicas destas lesões, uma vez que estas geralmente possuem tamanhos maiores em comparação às demais lesões.

A matriz de confusão se trata de uma tabela contendo os dados provenientes dos experimentos realizados com a abordagem. Com base nestes dados foi possível resumir as informações relacionadas ao desempenho da abordagem e comparar com os resultados obtidos com trabalhos similares no estado da arte. A Figura 43 apresenta a matriz de confusão obtida pela abordagem para detecção com otimizador SGD e *Tilling* durante a etapa de validação no conjunto de dados DDR. Convém destacar que a matriz de confusão resultante da detecção de objetos apresenta características diferentes quando comparada a problemas que envolvam somente a classificação de objetos em imagens pois a maioria dos erros do modelo são associados à classe de fundo (*background*) e não com as demais classes. Além disso, os resultados apresentados na matriz de confusão irão variar de acordo com o limite de confiança definido.

Na detecção de objetos é comum que sejam apresentadas na matriz de confusão informações referentes a falsos positivos (FP) e falsos negativos (FN) de fundo (*background*). Nesse sentido, o limite de confiança estabelecido para a detecção dos objetos presentes nestas imagens irá impactar diretamente nos resultados obtidos de *background* FP e *background* FN.

Portanto, a última linha da matriz de confusão se refere aos objetos do *Ground Truth* que não foram detectados pela abordagem (*background* FN) e, portanto, consideradas como fundo. Já a última coluna da matriz de confusão são as detecções realizadas pela abordagem que não têm nenhum rótulo correspondente no *Ground*

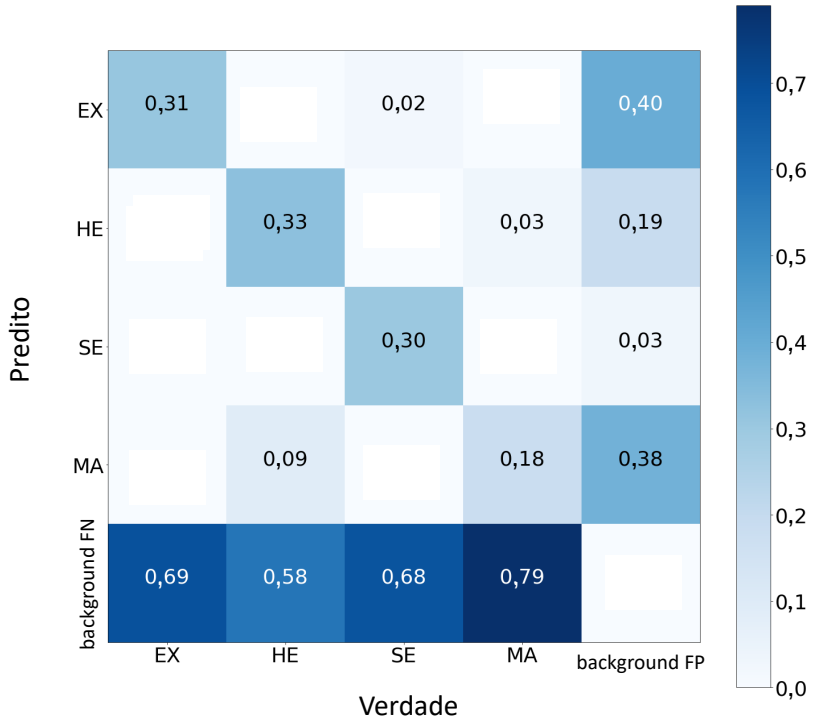


Figura 43 – Matriz de confusão obtida pela abordagem para detecção com otimizador SGD e *Tilling* durante a etapa de validação no conjunto de dados DDR.

Truth (*background* FP), ou seja, fundo da imagem detectado como lesão.

O limite de confiança é aplicado para filtrar as caixas delimitadoras de um possível objeto a fim de eliminar caixas delimitadoras com baixa pontuação de confiança por meio de um algoritmo de Supressão Não-Máxima, que desconsidera objetos detectados com IoU menor que o limite definido. Dessa forma, caso seja definido um limite de confiança alto, como por exemplo 0,90, haverá pouca confusão entre as classes e baixos resultados de *background* FP, porém haverá uma eliminação acentuada de lesões de fundo detectadas corretamente (embora com limite de confiança baixo), porém com limite de confiança inferior a 0,90. Em contrapartida, se for definido um limite de confiança de 0,25 haverá uma geração maior de *background* FP e *background* FN, uma vez que aumenta a probabilidade de o modelo detectar fundo como sendo lesão e vice-versa.

Portanto, à medida que o limite de confiança tender para 1 os FPs de fundo tenderão para 0. Os resultados apresentados na matriz de confusão foram calculados em limite fixo de confiança de 0,25, que está alinhado com a configuração padrão de inferência constante no arquivo `detect.py` da abordagem. Em síntese, com limites de confiança mais baixos irá melhorar os resultados de mAP, mas também irá produzir uma quantidade maior de *background* FP que constará na matriz de confusão, ao passo que se aumentar o limite de confiança haverá uma diminuição de *background* FP na matriz de confusão, porém, com prejuízo no mAP uma vez que mais lesões detectadas corretamente poderão ser desconsideradas.

As células com tonalidades mais escuras de azul indicam um número maior de amostras. A matriz de confusão apresenta os acertos do modelo na predição das lesões de fundo na diagonal principal, ao passo que os valores fora da diagonal principal correspondem a erros de predição.

É possível verificar que a maior incidência de *background* FN ocorreu nos Microaneurismas (com 79%), seguido por Exsudatos Duros (com 69%), Exsudatos Algodonosos (com 68%) e Hemorragias (com 58%). Quanto aos erros de *background* FP, a maior incidência ocorreu nos Exsudatos Duros (com 40%), seguido dos Microaneurismas (com 38%), Hemorragias (com 19%) e Exsudatos Algodonosos (com 3%). Também é possível observar que 9% de Hemorragias foram detectadas incorretamente como Microanerismas, 2% de Exsudatos Algodonosos foram detectados incorretamente como Exsudatos Duros e 3% de Microaneurismas foram detectados incorretamente como Hemorragias.

Dessa forma, os resultados apresentados na matriz de confusão não podem ser comparados diretamente com os resultados de AP apresentados neste trabalho, pois os valores estão associados à área sob a curva PR, conforme apresentado na Figura 42. Um bom mAP produzido por um limite de confiança baixo, por exemplo, necessariamente conterá milhares de FPs, empurrados para o canto inferior direito da curva PR, com tendências de Revocação para 1 e Precisão para 0, conforme pode ser observado na Figura 42.

Durante a etapa de teste da abordagem para detecção no conjunto de dados DDR utilizando o otimizador SGD, o melhor resultado obtido alcançou um mAP de 0,1430 (indicado em negrito), e valores de AP com limite de IoU de 0,5 iguais a 0,2100, 0,1380, 0,0530 e 0,1710, para Exsudatos Duros (EX), Exsudatos Algodonosos (SE), Microaneurismas (MA) e Hemorragias (HE), respectivamente, conforme apresentado na Tabela 13. Ambos resultados obtidos pela abordagem, com e sem a utilização de *Tilling*, alcançaram resultados superiores aos trabalhos relacionados, que também realizaram a detecção de lesões de fundo em imagens do conjunto de teste do conjunto de dados DDR.

Também realizaram-se experimentos durante a etapa de validação da abordagem na detecção no conjunto de dados DDR utilizando o otimizador Adam, em que o melhor resultado obtido foi utilizando o método de *Tilling*, com um mAP de 0,2630 (indicado em negrito), e valores de AP com limite de IoU de 0,5 iguais a 0,2240, 0,3650, 0,1110 e 0,3520, para Exsudatos Duros (EX), Exsudatos Algodonosos (SE), Microaneurismas (MA) e Hemorragias (HE), respectivamente, conforme apresentado na Tabela 14.

Tabela 13 – Resultados obtidos pela abordagem para detecção com otimizador SGD em comparação aos trabalhos relacionados com as métricas AP e mAP para o limite de IoU de 0,5 no conjunto de teste do conjunto de dados DDR.

Modelos	AP				mAP
	EX	**SE**	**MA**	**HE**	
SSD (LI et al., 2019)	0,0002	0	0,0001	0,0056	0,0015
YOLO (LI et al., 2019)	0,0012	0	0	0,0109	0,0030
YOLOv5-S (*baseline*) (JOCHER, 2020)	0,1270	0,0866	0,0340	0,1290	0,0942
Abordagem para Detecção+SGD sem *Tilling*	0,1430	0,2040	0,0280	0,1480	0,1310
Abordagem para Detecção+SGD com *Tilling*	0,2100	0,1380	0,0530	0,1710	**0,1430**

Tabela 14 – Resultados obtidos pela abordagem para detecção com otimizador Adam em comparação aos trabalhos relacionados com as métricas AP e mAP para o limite de IoU de 0,5 no conjunto de validação do conjunto de dados DDR.

Modelos	AP				mAP
	EX	**SE**	**MA**	**HE**	
SSD (LI et al., 2019)	0	0,0227	0	0,0007	0,0059
YOLO (LI et al., 2019)	0,0039	0	0	0,0101	0,0035
YOLOv3+Adam+*Dropout* (ALYOUBI; ABULKHAIR; SHALASH, 2021)	-	-	-	-	0,2160
YOLOv5-S (*baseline*) (JOCHER, 2020)	0,1360	0,2170	0,0759	0,2360	0,1662
Abordagem para Detecção+Adam sem *Tilling*	0,1640	0,4020	0,0610	0,3290	0,2390
Abordagem para Detecção+Adam com *Tilling*	0,2240	0,3650	0,1110	0,3520	**0,2630**

A utilização do otimizador Adam resultou em um mAP superior ao resultado obtido pela abordagem para detecção com otimizador SGD, apresentado na Tabela 12. A abordagem apresentou resultados superiores a todos os trabalhos de mesmo propósito encontrados na literatura.

A Figura 44 apresenta o gráfico da curva PR com limite de IoU de 0,5 obtido durante a etapa de validação utilizando a abordagem para detecção com otimizador Adam e *Tilling* no conjunto de dados DDR. No gráfico são plotados os valores de AP obtidos pela lesões de fundo, conforme os resultados apresentados na Tabela 14, cujo valor de *mean Average Precision* obtido por todas as classes corresponde a 0,263. Analisando a curva PR constata-se que utilizando o otimizador Adam a abordagem proposta apresentou resultados similares aos obtidos utilizando o otimizador SGD, ou seja houve uma taxa elevada de erros na detecção de Microaneurismas (MA) (curva na cor vermelha).

Os melhores resultados foram alcançados na predição dos Exsudatos Algodonosos (SE) (curva na cor em verde) e Hemorragias (HE) (curva na cor em laranja), respectivamente. É possível verificar na matriz de confusão da Figura 45 a alta taxa de *background* FN (89%) e alta taxa de *background* FP (34%) dos microaneurismas. A taxa de *background* FP dos exsudatos duros também se destaca das demais classes de lesões, alcançando 44%. As razões que ocasionaram as altas taxas de FN

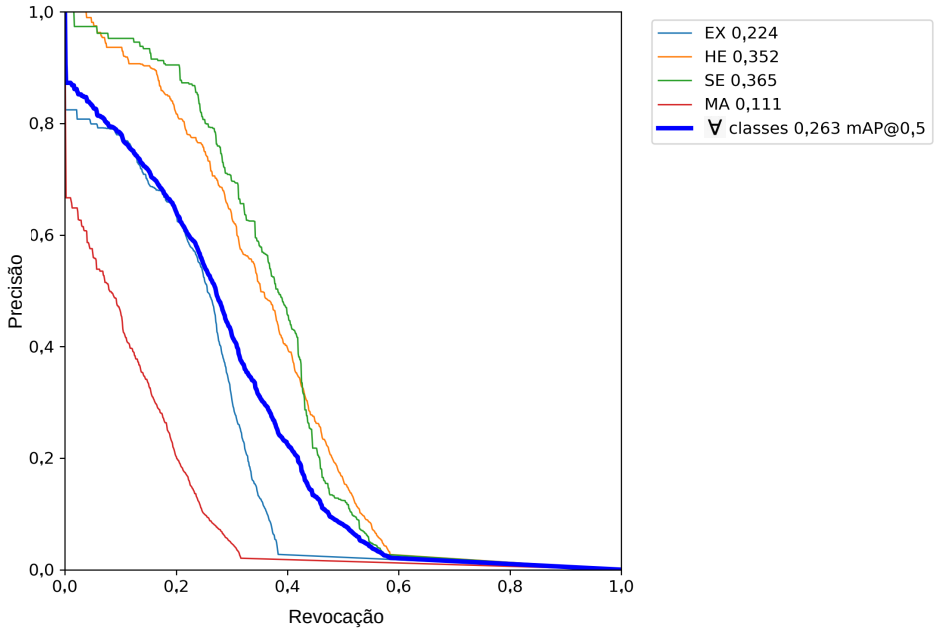


Figura 44 – Gráfico da curva PR com limite de IoU de 0,5 obtido durante a etapa de validação da abordagem para detecção com otimizador Adam e *Tilling* no conjunto de dados DDR.

e FP, tanto na detecção de microaneurismas quanto nos exsudatos duros, já foram discutidas anteriormente.

A Figura 45 apresenta a matriz de confusão obtida pela abordagem para detecção com otimizador Adam e *Tilling* durante a etapa de validação no conjunto de dados DDR. É possível verificar que a maior incidência de *background* FN ocorreu nos Microaneurismas (com 80%), seguido por Exsudatos Duros (com 67%), Exsudatos Algodonosos (com 63%) e Hemorragias (com 58%). Quanto aos erros de *background* FP, a maior incidência ocorreu nos Exsudatos Duros (com 44%), seguido dos Microaneurismas (com 34%), Hemorragias (com 17%) e Exsudatos Algodonosos (com 5%). Também é possível observar que 10% de Hemorragias foram detectadas incorretamente como Microaneurismas, 2% de Exsudatos Algodonosos foram detectados incorretamente como Exsudatos Duros e 2% de Microaneurismas foram detectados incorretamente como Hemorragias.

A Figura 46 apresenta um lote com imagens de fundo do conjunto de dados DDR juntamente com as anotações (*Ground Truth*) das lesões de fundo após as etapas de pré-processamento e aumento de dados que foram utilizadas para a validação da abordagem para detecção utilizando otimizador Adam e *Tilling*. Já na Figura 47 são apresentadas as detecções das lesões de fundo realizadas no mesmo lote de imagens de fundo descrito anteriormente.

A abordagem conseguiu detectar satisfatoriamente as lesões de fundo como pode ser observado o microaneurisma localizado na imagem “007-3038-100_3.jpg”, ou a hemorragia e exsudato algodonoso na imagem “007-6127-300_1.jpg”, ou ainda o microaneurisma e exsudato algodonoso da imagem “007-6121-300_3.jpg”, e microaneu-

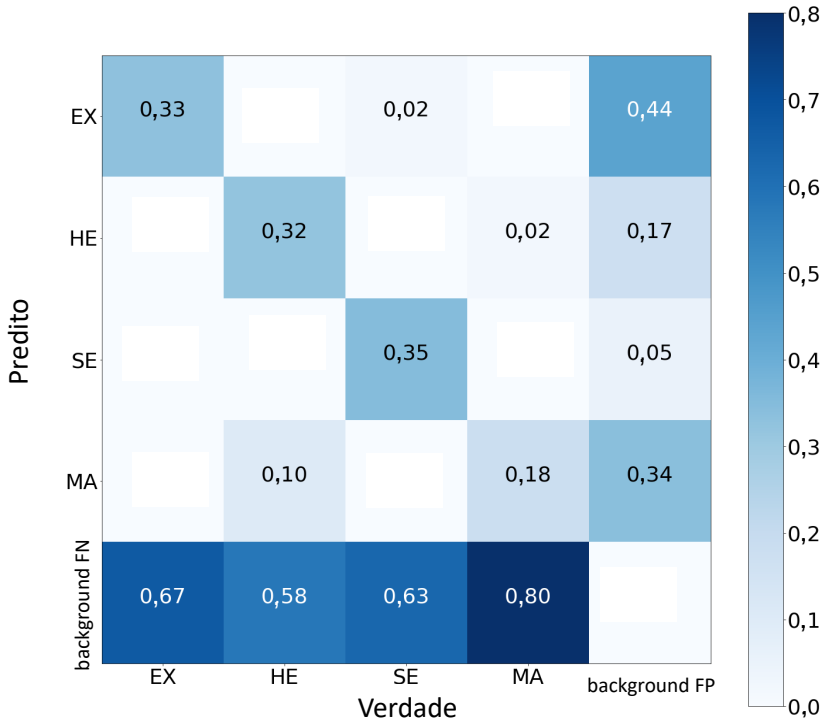


Figura 45 – Matriz de confusão obtida pela abordagem para detecção com otimizador Adam e *Tilling* durante a etapa de validação no conjunto de dados DDR.

rismas da imagem “007-6121-300_1.jpg”. Entretanto, ocorreram casos que a abordagem não conseguiu detectar as lesões, ou detectou erroneamente, como no caso de um dos microaneurismas não localizado na imagem “007-3045-100_3.jpg”, a hemorragia da imagem “007-6127-300_3.jpg”, e dois microaneurismas da imagem “007-3045-100_1.jpg”.

Ainda assim, também houve situações que a abordagem detectou objetos na imagem “007-3045-100_3.jpg” como sendo exsudatos duros, mesmo não havendo anotações destas lesões no *Ground Truth* do conjunto de dados DDR. Em função desta imagem apresentar microaneurismas no mesmo setor dos exsudatos duros localizados, existe a possibilidade destes exsudatos terem sido detectados corretamente, mesmo não tendo sido localizados originalmente no conjunto de dados. Porém, não há como ter a confirmação desta informação, uma vez que somente com a verificação da ausência de luminescência destas lesões, comprovada por meio de uma angiografia da imagem (não disponibilizada no conjunto de dados), seria possível ter certeza sobre o diagnóstico. De qualquer forma, com base nas lesões detectadas e a capacidade de generalização verificada após as predições realizadas em imagens não conhecidas *a priori*, a abordagem para detecção demonstrou ser uma importante ferramenta para o auxílio de diagnóstico médico.

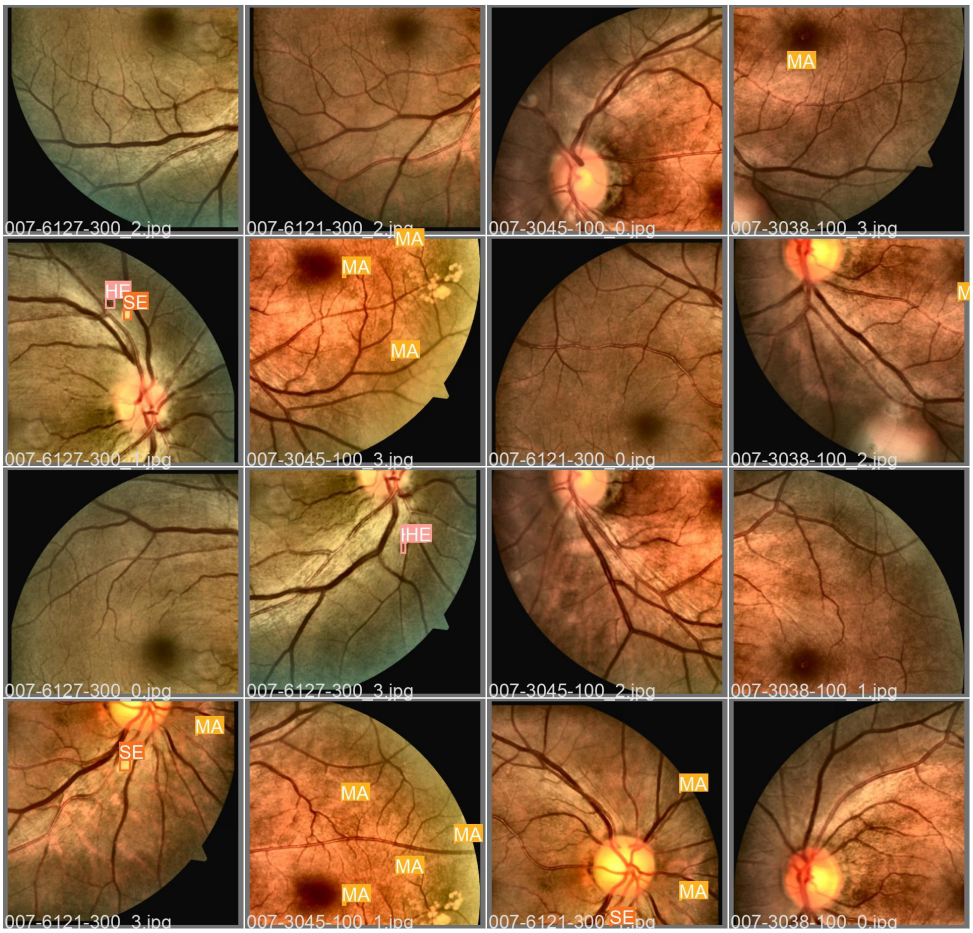


Figura 46 – Exemplo de lote com imagens de fundo do conjunto de dados DDR juntamente com as anotações (*Ground Truth*) das lesões de fundo após as etapas de pré-processamento e aumento de dados que foram utilizadas para a validação da abordagem para detecção.

Durante a etapa de teste da abordagem no conjunto de dados DDR utilizando o otimizador Adam, o melhor resultado obtido alcançou um mAP de 0,1540 (indicado em negrito), e valores de AP com limite de IoU de 0,5 iguais a 0,2210, 0,1570, 0,0553 e 0,1840, para Exsudatos Duros (EX), Exsudatos Algodonosos (SE), Microaneurismas (MA) e Hemorragias (HE), respectivamente, conforme apresentado na Tabela 15. Assim como no conjunto de validação, a abordagem para detecção (com e sem a utilização de *Tilling*), obteve resultados superiores aos trabalhos relacionados testados no conjunto de teste do conjunto de dados DDR.

É possível verificar que o método de otimização que apresentou os melhores resultados foi o Adam. Dessa forma, pode-se concluir que este otimizador tem grande potencial para aplicação em problemas envolvendo detecção de lesões de fundo. Porém, como trabalhos futuros pretende-se realizar experimentos com outros métodos de otimização no estado da arte utilizando diferentes variações de hiperparâmetros.

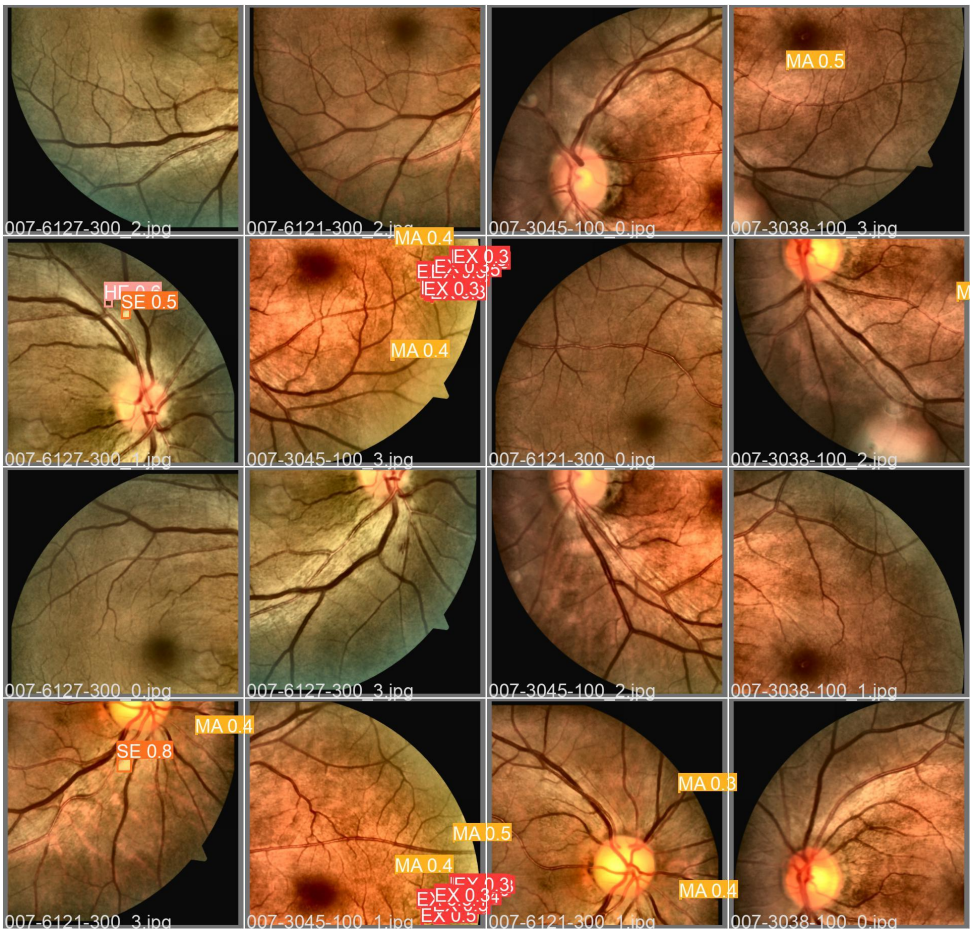


Figura 47 – Lote com imagens de fundo do conjunto de dados DDR com lesões de fundo detectadas pela abordagem para detecção durante a etapa de validação.

Tabela 15 – Resultados obtidos pela abordagem para detecção com otimizador Adam em comparação aos trabalhos relacionados com as métricas AP e mAP para o limite de IoU de 0,5 no conjunto de teste do conjunto de dados DDR.

Modelos	AP				mAP
	EX	**SE**	**MA**	**HE**	
SSD (LI et al., 2019)	0,0002	0	0,0001	0,0056	0,0015
YOLO (LI et al., 2019)	0,0012	0	0	0,0109	0,0030
YOLOv5-S (*baseline*) (JOCHER, 2020)	0,1410	0,0799	0,0344	0,1300	0,0963
Abordagem para Detecção+Adam sem *Tilling*	0,1540	0,2110	0,0296	0,1590	0,1380
Abordagem para Detecção+Adam com *Tilling*	0,2210	0,1570	0,0553	0,1840	**0,1540**

A seguir são apresentados os resultados obtidos com as métricas: Precisão, que considera entre todas as classificações positivas que o modelo fez, quantas estão corretas; a Revocação, que assume dentre todas as situações da classe positiva como valor esperado, quantas estão corretas; e, o F1-*score*, que calcula a média harmônica entre precisão e revocação. Os melhores resultados alcançados pela abordagem para detecção no conjunto de dados DDR foram obtidos com a utilização do otimizador Adam e o método de *Tilling*, conforme a métrica F1-*score* obtida nas etapas de

validação e teste, com valores de 0,3485 (indicado em negrito) e 0,2521 (indicado em negrito), respectivamente, conforme apresentado na Tabela 16.

Tabela 16 – Resultados obtidos com as métricas de Precisão, Revocação e F1-*score* com os otimizadores SGD e Adam durante as etapas de validação e teste utilizando o conjunto de dados DDR.

Modelos	**Precisão Validação**	**Revocação Validação**	**F1-*score* Validação**	**Precisão Teste**	**Revocação Teste**	**F1-*score* Teste**
Abordagem para Detecção+SGD sem *Tilling*	0,4533	0,2233	0,2992	0,3270	0,1540	0,2094
Abordagem para Detecção+Adam sem *Tilling*	0,4618	0,2484	0,3231	0,3060	0,1710	0,2194
Abordagem para Detecção+SGD com *Tilling*	0,4775	0,2653	0,3411	0,3390	0,1820	0,2368
Abordagem para Detecção+Adam com *Tilling*	0,4462	0,2859	**0,3485**	0,3410	0,2000	**0,2521**

Os valores de Precisão e Revocação foram calculados no valor de F1-*score* máximo, avaliado sobre 1.000 valores de confiança de 0 a 1 (diferentemente dos resultados apresentados na matriz de confusão em que o limite de confiança é fixado em 0,25 (ZHU et al., 2021)), onde o ponto máximo de F1-*score* é selecionado (melhor equilíbrio entre Precisão de valor mais alto e Revocação de valor mais alto). A Figura 48 ilustra o gráfico de F1-*score* obtido pela abordagem utilizando otimizador Adam com *Tilling* no conjunto de validação do conjunto de dados DDR. No gráfico é possível verificar que o F1 máximo obtido para as classes de lesões preditas foi de 0,35 para um limite de confiança de 0,240.

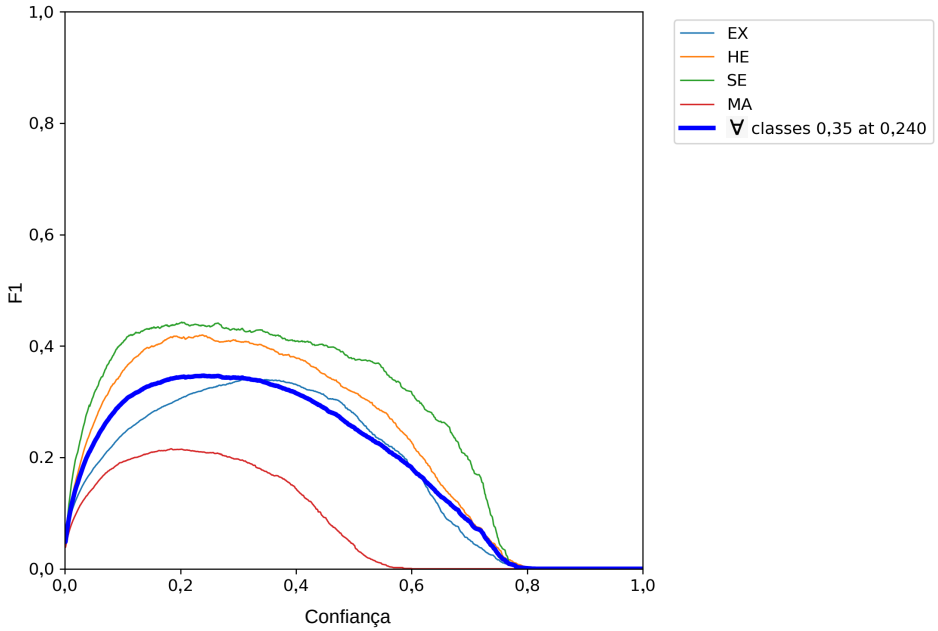


Figura 48 – Gráfico do F1-*score* obtido pela abordagem para detecção com otimizador Adam e *Tilling* durante a etapa de validação no conjunto de dados DDR.

O tempo médio de inferência para detectar as lesões de fundo no conjunto de dados DDR nas etapas de validação e teste da abordagem é apresentado na Tabela 17. A abordagem para detecção sem *Tilling* teve o menor tempo médio de inferência por imagem com otimizador Adam, com 14,1 ms, ao passo que com *Tilling* o menor tempo médio de inferência foi alcançado com otimizador SGD, com 4,6 ms (indicado em negrito). Porém, o destaque fica para o tempo de inferência da abordagem com a utilização de *Tilling*, que chegou a ser em torno de 3 vezes mais rápido em comparação ao tempo de inferência da abordagem aplicada nas imagens sem a realização de *Tilling*. Portanto, além de aumentar a precisão da abordagem na detecção das lesões de fundo, o método de *Tilling* tornou o processo de predição mais rápido.

Tabela 17 – Tempo médio de inferência para detectar as lesões de fundo no conjunto de dados DDR nas etapas de validação e teste da abordagem para detecção.

Modelos	Tempo de Inferência (ms)	
	Validação	**Teste**
Abordagem para Detecção+SGD sem *Tilling*	15,7	13,0
Abordagem para Detecção+Adam sem *Tilling*	14,1	21,1
Abordagem para Detecção+SGD com *Tilling*	**4,6**	**5,9**
Abordagem para Detecção+Adam com *Tilling*	5,5	7,5

Com o propósito de avaliar a precisão da abordagem em diferentes conjuntos de dados públicos de RD realizaram-se também experimentos com o conjunto de imagens de Retinopatia Diabética IDRiD (PORWAL et al., 2020). Durante a etapa de validação no conjunto de dados IDRiD, o melhor resultado obtido pela abordagem foi utilizando o otimizador SGD com o método de *Tilling*, com um mAP de 0,3280 (indicado em negrito), e valores de AP com limite de IoU de 0,5 iguais a 0,2630, 0,5340, 0,2170 e 0,2980, para Exsudatos Duros (EX), Exsudatos Algodonosos (SE), Microaneurismas (MA) e Hemorragias (HE), respectivamente, conforme apresentado na Tabela 18.

Tabela 18 – Resultados obtidos pela abordagem com *Tilling* e os otimizadores SGD e Adam com as métricas AP e mAP para o limite de IoU de 0,5 no conjunto de validação do conjunto de dados IDRiD.

Modelos	AP				mAP
	EX	**SE**	**MA**	**HE**	
Abordagem para Detecção+SGD sem *Tilling*	0,1030	0,2940	0,0601	0,2460	0,1760
Abordagem para Detecção+Adam sem *Tilling*	0,1040	0,1810	0,0723	0,1350	0,1230
Abordagem para Detecção+SGD com *Tilling*	0,2630	0,5340	0,2170	0,2980	**0,3280**
Abordagem para Detecção+Adam com *Tilling*	0,2670	0,2740	0,2100	0,3200	0,2680

O melhor resultado de mAP obtido pela abordagem para detecção durante a validação utilizando o conjunto de dados IDRiD superou os resultados obtidos no conjunto de dados DDR, atestando portanto a capacidade de generalização do método adotado pela abordagem para a detecção das lesões de fundo.

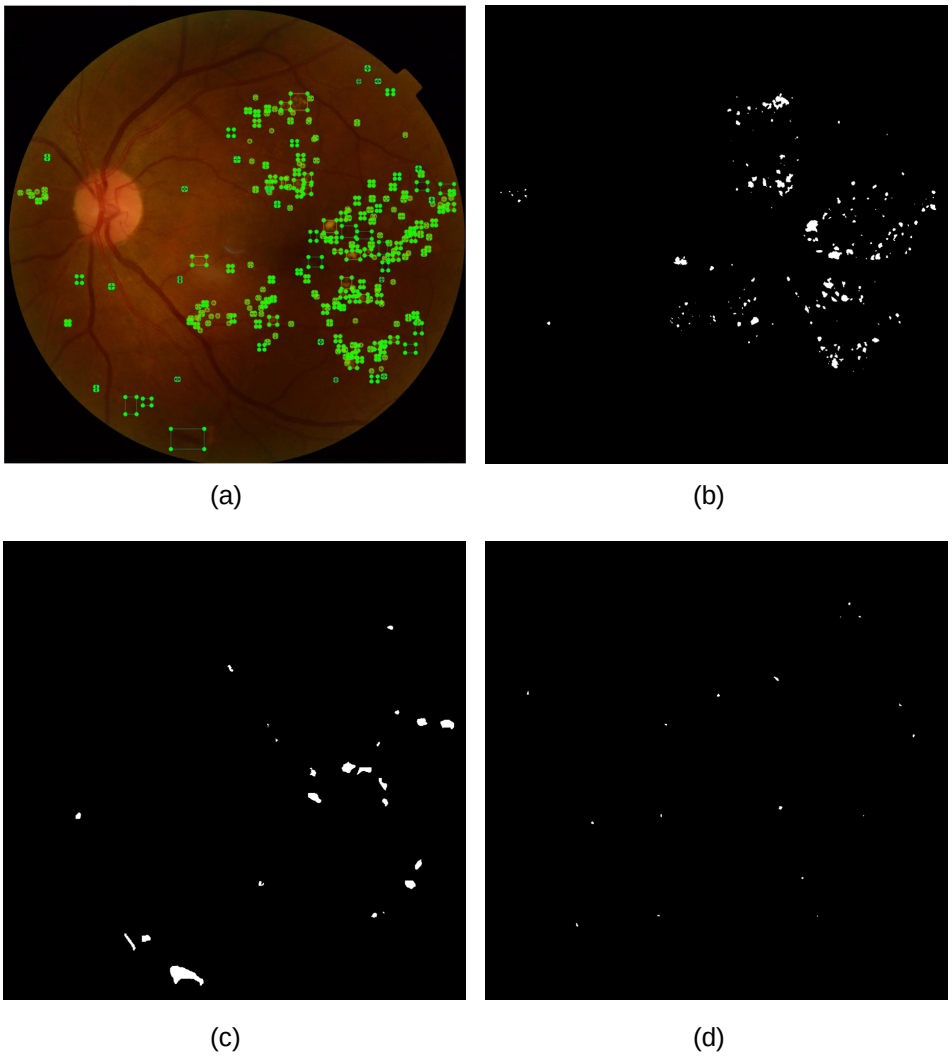
Na etapa de teste da abordagem utilizando o conjunto de dados IDRiD, o melhor resultado foi obtido com otimizador Adam e a utilização de *Tilling*, tendo alcançado um mAP de 0,2950 (indicado em negrito), e valores de AP com limite de IoU de 0,5 iguais a 0,2530, 0,4090, 0,2210 e 0,2970, para Exsudatos Duros (EX), Exsudatos Algodonosos (SE), Microaneurismas (MA) e Hemorragias (HE), respectivamente, conforme apresentado na Tabela 19.

Tabela 19 – Resultados obtidos pela abordagem para detecção com *Tilling* e os otimizadores SGD e Adam com as métricas AP e mAP para o limite de IoU de 0,5 no conjunto de teste do conjunto de dados IDRiD.

Modelos	AP				mAP
	EX	**SE**	**MA**	**HE**	
Abordagem para Detecção+SGD sem *Tilling*	0,1260	0,3000	0,0787	0,2630	0,1920
Abordagem para Detecção+Adam sem *Tilling*	0,0993	0,2640	0,0661	0,1380	0,1420
Abordagem para Detecção+SGD com *Tilling*	0,2390	0,3940	0,2010	0,2890	0,2810
Abordagem para Detecção+Adam com *Tilling*	0,2530	0,4090	0,2210	0,2970	**0,2950**

A Figura 49(a) corresponde à imagem de fundo “007-3711-200.jpg” do conjunto de teste do conjunto de dados DDR, juntamente com as anotações (*Ground Truth*) das lesões de fundo; a Figura 49(b) é a máscara de segmentação dos exsudatos duros; a Figura 49(c) é a máscara de segmentação das hemorragias; e, a Figura 49(d) é a máscara de segmentação dos microaneurismas. Importante destacar que não há a presença de exsudatos algodonosos nesta imagem de fundo.

A Figura 50 demonstra a detecção de lesões de fundo realizada pela abordagem para detecção e o percentual de confiança obtido em cada objeto localizado na imagem de fundo “007-3711-200.jpg” do conjunto de teste do conjunto de dados DDR. O *Ground Truth* desta imagem de fundo é apresentado na Figura 49. Esta imagem da retina tem um fundo mais escuro, característica recorrente em diversas imagens de fundo dos conjuntos de dados públicos investigados, que ocasiona frequentemente problemas na detecção das lesões (principalmente microaneurismas e hemorragias), gerando altas taxas de erros durante a classificação, em que uma lesão é considerada erroneamente como fundo (*background* FN) ou, o contrário, em que o fundo é considerado erroneamente como uma lesão (*background* FP). Para estes casos, as técnicas de processamento de imagens aplicadas no bloco de pré-processamento da abordagem têm um papel importante, pois objetivam minimizar estes problemas por meio da redução de ruídos e melhoria do contraste destas imagens, por exemplo.



(a) (b)

(c) (d)

Figura 49 – Exemplo de imagem de fundo do conjunto de dados acompanhada das máscaras de segmentação das lesões presentes na imagem. Em (a), a imagem de fundo “007-3711-200.jpg” do conjunto de teste do conjunto de dados DDR, juntamente com as anotações (*Ground Truth*) das lesões de fundo; (b) máscara de segmentação dos exsudatos duros; (c) máscara de segmentação das hemorragias; e (d) máscaras de segmentação dos microaneurismas.

Outro aspecto que também pode ser verificado nesta imagem de fundo são as microlesões identificadas, como no caso dos microaneurismas. São lesões extremamente pequenas que acabam dificultando sua detecção. Para estes casos, destaca-se mais uma vez a importância, por exemplo, da aplicação do método de *Tilling* (bloco de pré-processamento da abordagem), da aplicação das transformações geométricas (bloco de aumento dados da abordagem) e das características arquiteturais da rede neural da abordagem, como por exemplo, a utilização de CSPs integrando *Backbone* e *Neck*, que visam minimizar problemas de dissipação de gradiente destas microlesões durante o treinamento da rede neural. Ainda sim, com todas estas características observadas nesta imagem de fundo que dificultam a identificação das lesões, é possível verificar que a abordagem para detecção realizou com precisão a detecção da

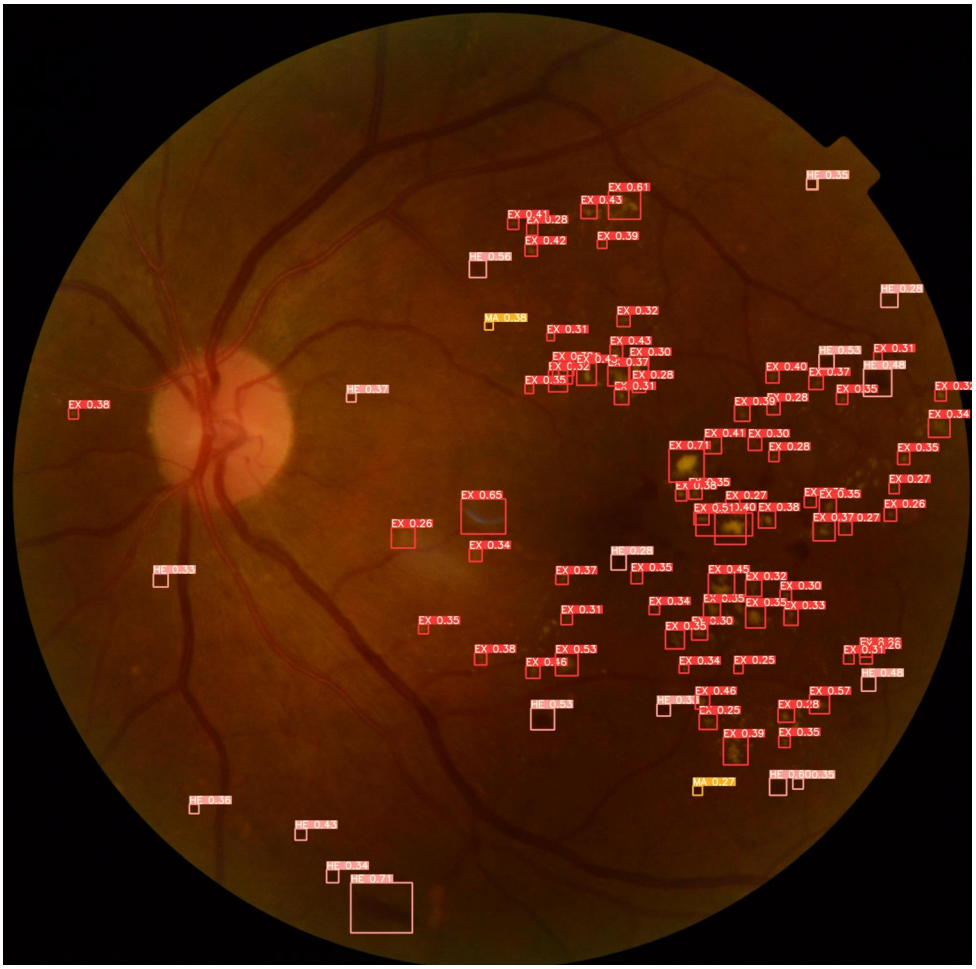


Figura 50 – Detecção de lesões de fundo realizada pela abordagem para detecção e o percentual de confiança obtido em cada objeto localizado na imagem de fundo “007-3711-200.jpg” do conjunto de teste do conjunto de dados DDR.

maioria das lesões de fundo, localizando exsudatos duros (EX), hemorragias (HE) e microaneurismas (MA) presentes na imagem.

A Figura 51 apresenta as mesmas detecções realizadas na imagem de fundo “007-3711-200.jpg” descritas anteriormente, porém com maior nível de detalhe para as lesões detectadas em torno da Mácula (MAIA, 2018), uma região posterior e central da retina que apresenta aspecto ovalado e pode ser observada como uma mancha redonda mais escura, cujo centro é conhecido como fóvea, uma região do olho humano responsável pela nitidez da visão (BESENCZI; TÓTH; HAJDU, 2016).

A Figura 52 apresenta a detecção de lesões de fundo na imagem “007-3892-200.jpg” do conjunto de teste do conjunto de dados DDR, em que é possível observar com maior nível de detalhe os diferentes aspectos morfológicos de algumas lesões identificadas, tais como os exsudatos duros, que aparecem aglomerados na forma de cacho na região central da imagem ou isolados em outras regiões da retina. As hemorragias também surgem sob diferentes formas e tamanhos, em diversos quadrantes da retina e, que em função destas características, frequentemente produzem erros de

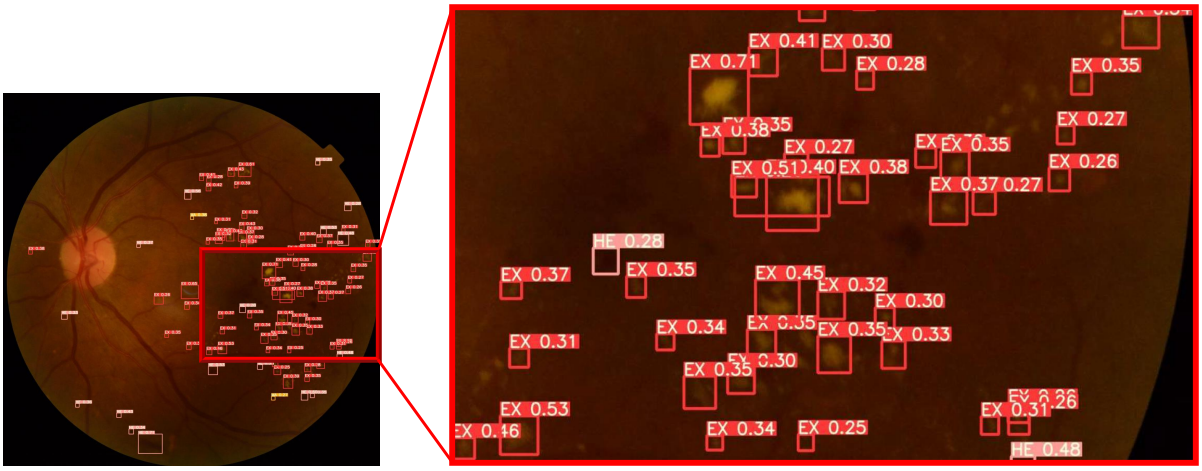


Figura 51 – Lesões detectadas na imagem de fundo “007-3711-200.jpg” em torno da mácula, região localizada no centro da retina que pode ser observada como uma mancha redonda mais escura, cujo centro é conhecido como fóvea, que tem a função de garantir o detalhamento das imagens formadas no campo de visão.

classificação devido a similaridades com os microaneurismas, conforme resultados apresentados nas matrizes de confusão dos experimentos (vide Figuras 43 e 45).

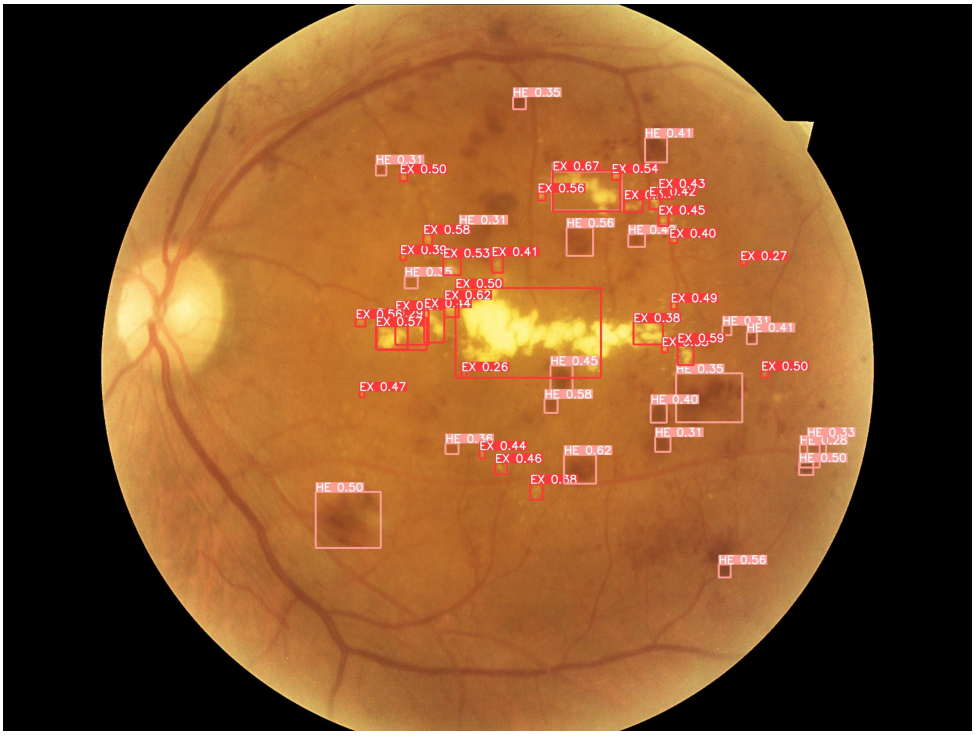


Figura 52 – Detecção de lesões de fundo na imagem “007-3892-200.jpg” do conjunto de teste do conjunto de dados DDR, na qual é possível observar diferentes aspectos morfológicos das lesões identificadas, como no caso dos exsudatos duros na região central da imagem e distribuídos em outras regiões da retina, ou das hemorragias, que assim como os exsudatos duros detectados também assumem diferentes formas e tamanhos, além de poderem se manifestar em diferentes regiões da retina.

Os trabalhos propostos por Alyoubi; Abulkhair; Shalash (2021) e Dai et al. (2021) apresentaram resultados no conjunto de validação. Diferentemente destes trabalhos, nesta abordagem adotou-se a mesma metodologia do trabalho proposto por Li et al. (2019), em que realizou-se a avaliação por meio da análise dos resultados obtidos tanto na etapa de validação quanto na etapa de teste utilizando o conjunto público de dados de Retinopatia Diabética DDR. Foi adotado este método com o propósito de evitar que a avaliação da abordagem fosse realizada somente no conjunto de validação, uma vez que este tipo de avaliação poderia ocasionar uma falsa impressão de que a abordagem é precisa quanto à detecção das lesões de fundo.

Realizar a avaliação da abordagem em um conjunto de validação (onde o modelo de rede neural é ajustado) e, em seguida, em um conjunto de teste (onde os dados não são conhecidos *a priori*) permitiu que a capacidade de generalização da abordagem fosse verificada adequadamente, sem o risco de vieses produzidos por um possível sobreajuste do modelo durante a validação. Afim de validar a capacidade preditiva da abordagem quanto à detecção de lesões de fundo também foi feita a avaliação no conjunto de dados IDRiD, em que se alcançaram resultados equivalentes aos obtidos no conjunto de dados DDR.

O trabalho de Dai et al. (2021) não foi comparado com a abordagem proposta para detecção, pois os autores utilizaram uma arquitetura de 2 estágios ao passo que aqui foi utilizada uma arquitetura de estágio único. Os autores não apresentaram os resultados de AP ou mAP, diferentemente dos demais trabalhos com propósito similar encontrados na literatura, o que impossibilita uma comparação adequada. Além disso, os autores utilizaram um conjunto de dados de RD privado para realizar o treinamento dos modelos de aprendizado profundo, o que dificulta a reprodutibilidade dos resultados obtidos utilizando o mesmo método.

No trabalho proposto por Li et al. (2019), os melhores resultados obtidos quanto à detecção de lesões de fundo no conjunto de dados DDR, usando modelos de estágio único, foram 0,0059 de mAP na etapa de validação com SSD, e 0,0030 na etapa de teste com YOLO. Santos et al. (2021), utilizando o conjunto de dados DDR, obteve na etapa de validação um mAP de 0,0716 com o modelo YOLOv4. Já no trabalho apresentado por Alyoubi; Abulkhair; Shalash (2021), o melhor resultado obtido pelos autores na etapa de validação com o conjunto de dados DDR foi um mAP de 0,1710, utilizando o modelo YOLOv3. Na abordagem proposta neste trabalho, implementada com base no modelo YOLOv5, obteve-se um mAP de 0,2630 na etapa validação e 0,1540 na etapa de teste, ambos resultados obtidos no conjunto de dados DDR.

Com um limite de confiança estabelecido em 0,25, identificaram-se as lesões com seus respectivos percentuais de confiança. Os resultados experimentais mostraram que a abordagem para detecção obteve maior precisão em comparação a trabalhos similares encontrados na literatura. Outro aspecto observado durante os experimen-

tos é que a abordagem proposta para detecção obteve maior precisão na detecção de Exsudatos Algodonosos, Hemorragias e Exsudatos Duros e, em contraste, uma precisão mais baixa na detecção de Microaneurismas.

A detecção dessas lesões por meio de sistemas informatizados é um desafio por inúmeros fatores, dentre os quais: as características de tamanho e formato destas lesões; ruídos e contraste das imagens disponíveis nos conjuntos de dados públicos de RD; a quantidade de exemplos anotados destas lesões, disponíveis em conjuntos de dados públicos de RD; e, a dificuldade de algoritmos de aprendizado profundo em detectar objetos muito pequenos, etc. Estes problemas foram relatados na literatura e também observados durante os experimentos que foram realizados. Assim, para contornar alguns destes problemas, propõe-se uma nova abordagem baseada em técnicas de processamento de imagens e uma arquitetura de rede neural profunda *Single-Stage* de última geração para detectar as lesões em imagens de fundo.

Para o problema relacionado à forma e tamanho dos objetos, em que lesões muito pequenas como os microaneurismas apresentam maior dificuldade de serem detectadas, aplicaram-se técnicas para aumentar o campo receptivo destas lesões, tais como o *Cropping* parcial do fundo preto das imagens e o *Tilling* das imagens de entrada para treinamento da rede neural. Também foi aplicada a técnica de aumento de dados baseada na transformação geométrica *Scale*, onde um zoom de 50% é realizado aleatoriamente nas imagens de entrada para que sejam criadas artificialmente novas imagens para treinamento e a rede neural possa extrair mais características, tornando-a mais eficiente na detecção das microlesões.

Desenvolveu-se um bloco de pré-processamento, em que primeiramente realizou-se uma filtragem nas imagens, a fim de remover *outliers* provenientes da captura destas imagens e, em seguida, aplicou-se o método de equalização adaptativa de histograma limitada por contraste para aumentar o contraste local das imagens de fundo e melhorar o realce das lesões.

Para minimizar o problema da pequena quantidade de exemplos de lesões anotadas nos conjuntos de dados públicos de RD, desenvolveu-se um bloco responsável pelo aumento de dados. Nesse bloco, foram aplicados diferentes métodos no estado da arte, tais como *Mosaic*, *MixUp*, *Copy-Paste* e transformações geométricas (*flipping*, *scaling*, *perspective*, *translation* e *shearing*). O propósito desta etapa foi criar artificialmente uma quantidade maior de imagens de exemplo com lesões anotadas para treinamento da abordagem, para permitir que a rede neural profunda extraísse uma quantidade maior de características das lesões e, consequentemente, que aumentasse a capacidade de generalização da abordagem proposta.

Em comparação aos trabalhos similares encontrados na literatura, é importante destacar as contribuições da abordagem proposta para detecção relacionadas à estrutura da rede neural profunda utilizada, tais como o uso de módulos CSP (C3) no *Backbone* e *Neck* da arquitetura, que minimizou problemas de dissipação de gradiente, causados pela quantidade de camadas densas. Além disso, por meio destes módulos, houve uma melhoria na velocidade de inferência e precisão na detecção das lesões, além de redução do custo computacional e uso de memória. Outra inovação na estrutura da abordagem proposta para detecção foi a utilização da função de ativação SiLU em toda a rede neural, com o intuito de simplificar a arquitetura e diminuir a quantidade de hiperparâmetros.

Aplicou-se o método *Threshold-Moving* durante o treinamento da rede neural, de modo que as amostras de imagens foram ponderadas por meio de uma métrica de precisão, com o objetivo de minimizar o desiquilíbrio da quantidade de exemplos das diferentes classes de lesões investigadas e evitar vieses de classificação em classes majoritárias. Por fim, realizaram-se os ajustes e testes da abordagem por meio de diferentes conjuntos de dados públicos de Retinopatia Diabética, dividindo estes conjuntos de dados em treino, validação e teste, a fim de realizar a avaliação da abordagem de acordo com os resultados obtidos nas diferentes métricas de desempenho adotadas.

6.4 Considerações sobre o Capítulo

Este Capítulo apresentou uma abordagem baseada em um modelo de rede neural profunda que detecta lesões de fundo associadas à Retinopatia Diabética. Foram mostradas as etapas de pré-processamento e aumento de dados, assim como a constituição da arquitetura de rede neural profunda utilizada.

Particionou-se os conjuntos de dados públicos de Retinopatia Diabética utilizados nos experimentos em uma proporção de 50:20:30 para aplicar o processo de treinamento, validação e teste, respectivamente. Usou-se uma arquitetura de rede neural profunda baseada na estrutura de uma YOLO de última geração, cuja implementação foi realizada por meio da biblioteca PyTorch.

Além disso, foram apresentados os cenários experimentais, testes e avaliação dos resultados obtidos pela abordagem proposta para detecção. Por meio da avaliação realizada verificou-se que os resultados foram promissores, superando os trabalhos relacionados encontrados na literatura, e demonstrando que a abordagem proposta para detecção foi eficaz na detecção das lesões da retina investigadas neste trabalho.

No próximo Capítulo, é apresentada a abordagem para a segmentação de instância de lesões em imagens de fundo, como uma alternativa à abordagem proposta para detecção de lesões de fundo apresentada neste Capítulo.

7 ABORDAGEM PARA SEGMENTAÇÃO DE INSTÂNCIA DE LESÕES DE FUNDO

Este Capítulo tem por objetivo apresentar a abordagem proposta para a segmentação de instância de lesões de fundo. Serão mostradas as etapas de pré-processamento e aumento de dados, assim como a constituição da arquitetura da rede neural profunda utilizada. Também são apresentados os cenários experimentais, tecnologias utilizadas, testes e avaliação dos resultados obtidos pela abordagem.

7.1 Materiais, Técnicas e Métodos

A abordagem foi desenvolvida com base na arquitetura Mask R-CNN (HE et al., 2020), conforme ilustrado no diagrama de blocos apresentado na Figura 53. Para a realização dos experimentos foi utilizado um equipamento com um Core i7 com 32 GB de RAM e uma GPU NVIDIA Titan Xp com 12 GB de VRAM. Para a construção da arquitetura utilizou-se a biblioteca de código aberto Detectron2 (WU et al., 2019; HONG et al., 2021; AMERIKANOS; MAGLOGIANNIS, 2022).

A arquitetura Mask R-CNN é um modelo capaz de detectar e segmentar instâncias de objetos. Este modelo estende a arquitetura de detecção de objetos Faster R-CNN (HE et al., 2020), adicionando uma estrutura paralela para prever as máscaras de segmentação dos objetos. A segmentação de instância combina tarefas de detecção de objetos, onde o objetivo é localizar e classificar objetos individualmente usando uma caixa delimitadora e, também, localizar cada *pixel* de cada objeto detectado na imagem.

Esta arquitetura funciona em dois estágios. O primeiro estágio consiste em usar uma Rede de Proposta de Região (RPN) (REN et al., 2017; ZHAO et al., 2019) para propor as caixas delimitadoras (em inglês, *Bouding Box* – BBox) de objetos candidatos. Já o segundo estágio tem por objetivo classificar as caixas candidatas, refinar as caixas e prever as máscaras dos objetos (Mask). Modelos que realizam a detecção de objetos, tais como Faster R-CNN (REN et al., 2017), SSD (KONISHI et al., 2016) e YOLO (REDMON et al., 2016), desenham uma caixa delimitadora ao redor dos ob-

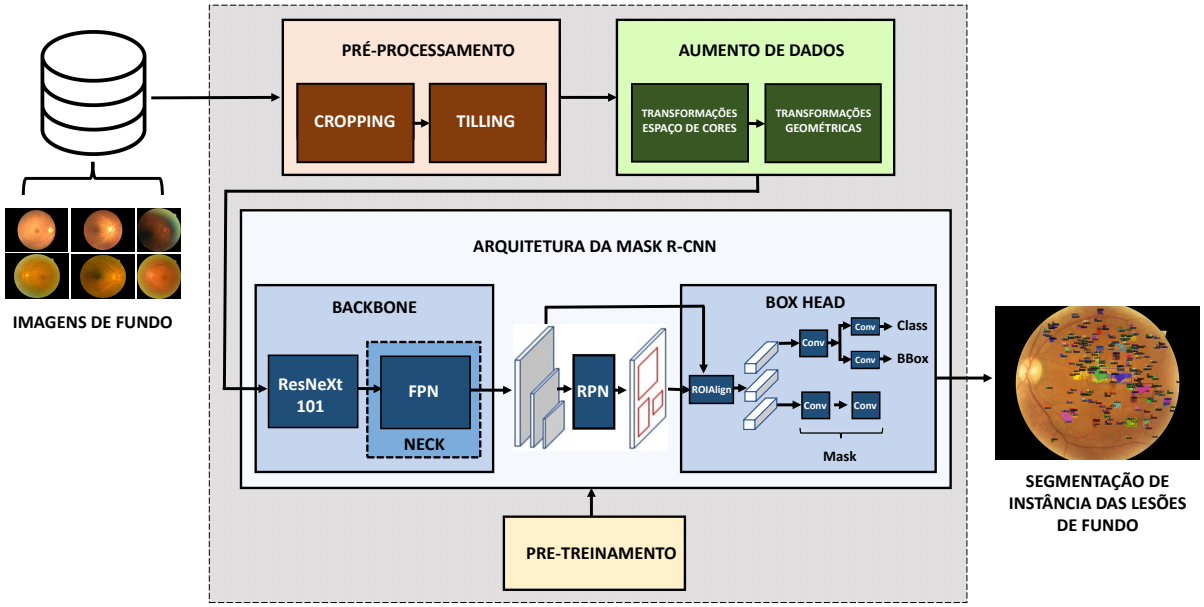


Figura 53 – Diagrama de blocos da abordagem para segmentação de instância de lesões de fundo. Primeiramente, as imagens são repassadas para o bloco de Pré-processamento para eliminação parcial do fundo preto das imagens (*Cropping*) e criação de sub-blocos das imagens (*Tilling*). Em seguida, as imagens pré-processadas são transferidas para o bloco de Aumento de Dados, no qual são criadas artificialmente novas imagens que serão utilizadas na camada de entrada do *Backbone* da arquitetura da rede neural para treinamento da abordagem. Entretanto, antes disso é realizada uma etapa de pré-treinamento com os pesos ajustados no conjunto de dados COCO.

jetos detectados, ao passo que na segmentação de instância também são fornecidas as máscaras em *pixels* para cada objeto localizado na imagem.

No caso da detecção de objetos há a possibilidade de lesões ficarem com suas caixas delimitadoras detectadas sobrepostas, dificultando a visualização destas lesões e, consequentemente, o diagnóstico. Com a segmentação de instância é possível detectar as lesões e também saber os locais dos *pixels* de cada lesão (bordas), portanto, sendo possível verificar o tamanho e a extensão de cada lesão detectada.

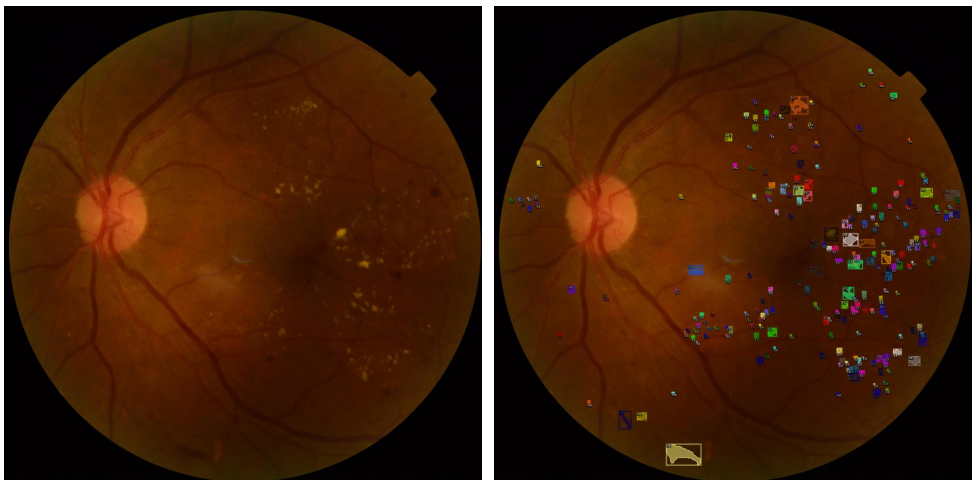
Além disso, os detectores de dois estágios frequentemente apresentam maior precisão na localização e classificação de objetos pequenos em relação a detectores de estágio único (JIAO et al., 2019; KIM; SUNG; PARK, 2020), principalmente quando estes objetos aparecem agrupados na imagem (MURTHY et al., 2020). Em detectores de dois estágios, há um estágio para a identificação de um subconjunto de regiões em uma imagem que pode conter um objeto e um segundo estágio para realizar a classificação do objeto em cada região. Este processo faz com que detectores de dois estágios sejam bons candidatos para realizar a detecção de objetos pequenos em imagens.

Com a abordagem para a segmentação de instância foi possível realizar, com elevado grau de precisão, a detecção das lesões de fundo, além de proporcionar uma melhor visualização da extensão de cada lesão detectada por meio da apresentação de sua máscara de segmentação. As informações fornecidas pela abordagem visam fornecer ao profissional de saúde um maior detalhamento sobre as lesões presentes em imagens da retina a fim de promover um diagnóstico precoce e mais assertivo da doença. A seguir, cada parte da metodologia adotada é detalhada.

7.1.1 Conjunto de Dados e Pré-processamento das Imagens

Para realização dos experimentos foram utilizados os conjuntos de dados públicos DDR e IDRiD. Foi utilizado o formato de anotações MS COCO[1], no qual as anotações dos objetos na forma de caixas delimitadoras e polígonos são armazenados em um arquivo no formato *JavaScript Object Notation* (JSON).

Para realizar o processo de criação das anotações das lesões na forma de polígonos, os arquivos das imagens juntamente com as máscaras binárias das lesões disponibilizados pelo conjunto de dados DDR foram utilizados para capturar o contorno das lesões por meio da função `find_contours()` do OpenCV. Após identificar o contorno das lesões, as anotações foram criadas com o auxílio da função `create_annotation_format()`. Por fim, estas anotações são transformadas para o formato padrão COCO JSON com a utilização da função `get_coco_json_format()`, a fim de serem utilizadas no treinamento da abordagem. A Figura 54 apresenta (a) um exemplo de imagem de fundo do conjunto de dados DDR e (b) a mesma imagem com as anotações das lesões no formato de caixas delimitadoras e as anotações geradas no formato de polígonos.



(a) **(b)**

Figura 54 – Imagens de fundo (a) do conjunto de dados DDR e (b) das anotações no formato de caixas delimitadoras e polígonos para treinamento da rede neural profunda.

[1] https://cocodataset.org/

Assim como os trabalhos de Santos et al. (2022) e Alyoubi; Abulkhair; Shalash (2021), a abordagem inclui uma etapa de pré-processamento para *cropping* parcial do plano de fundo preto das imagens da retina. A remoção do plano de fundo preto da imagem de fundo diminui a geração de falsos positivos durante a detecção das lesões, pois apenas os *pixels* da retina possuem informações significativas e o restante é considerado plano de fundo.

Para realizar a operação de remoção parcial do fundo preto foi utilizada a Transformada de Hough (HT) (NIXON; AGUADO, 2020). Porém, antes de identificar o contorno da retina foi realizado um pré-processamento com Filtro de Mediana, a fim de suavizar as imagens e eliminar detalhes irrelevantes para a detecção da circunferência da retina. Quando a forma da retina é encontrada na imagem de fundo, então o mapeamento de todos seus pontos no espaço de parâmetros se agrupam em torno dos valores de parâmetros que correspondem à sua forma. Cabe ressaltar que antes de aplicar a HT foi necessário fazer a limiarização das imagens, seguida pela detecção das bordas com o Filtro de Sobel. Após a localização da retina foi possível transformar sua circunferência em seu retângulo equivalente, para então ser realizada a remoção parcial do fundo preto da imagem.

A última etapa de pré-processamento é a realização de *Tilling*. Assim como o trabalho proposto por Santos et al. (2022), a abordagem para segmentação de instância inclui a operação de *Tilling*, em que as imagens originais são recortadas em blocos (*tiles*). Então, os blocos resultantes desta operação são utilizados na camada de entrada da rede neural profunda para treinamento do modelo. Com este método, as imagens são divididas em sub-imagens e o tamanho das sub-imagens resultantes varia de acordo com o tamanho original de cada imagem do conjunto de dados. As imagens foram particionadas em *tiles* de 2×2. Cada sub-imagem gerada neste processo permaneceu com suas respectivas lesões e anotações (*Ground Truth*) e, portanto, não havendo perda de informações.

Entretanto, quando o particionamento das imagens é realizado para criação dos *tiles* pode ocorrer de lesões estarem presentes no local de recorte destes blocos. Para minimizar o risco de perda de informações nestes casos foi definida uma área de *overlap*, em que cada bloco tem uma área de sobreposição de 15% com seus blocos vizinhos. Desse modo, não são perdidas informações de lesões que estejam presentes no local de recorte dos blocos, uma vez que estas informações são replicadas em diferentes blocos. Após a aplicação de *Tilling*, conforme será demonstrado na Seção 7.3, houve um aumento na precisão da abordagem na detecção das lesões em razão de mais informações destas lesões serem extraídas durante a etapa de treinamento da rede neural.

7.1.2 Aumento de Dados

Foi realizado o método de aumento de dados *on-the-fly*, no qual o carregador de dados da Detectron2 realiza a criação de imagens artificiais da retina com base no conjunto de dados público de RD usado para treinar a abordagem. O objetivo foi gerar mais dados de treinamento para que a rede neural profunda utilizada no *Backbone* pudesse extrair mais características das lesões de fundo. O aumento de dados foi realizado com diferentes tipos de dados, ou seja, com as imagens de fundo e as anotações das lesões. A seguir, são explicadas as técnicas utilizadas na etapa de aumento dos dados da abordagem.

Uma sequência estática foi definida para a aplicação do aumento de dados das imagens de fundo utilizadas durante o treinamento da rede neural profunda. Os métodos de aumento utilizados transformaram as imagens de forma que as anotações das lesões também fossem preservadas. Foram utilizados dois tipos diferentes de aumentos: *Color Space Transformations* e *Geometric Transformations* (SHORTEN; KHOSHGOFTAAR, 2019). Na categoria de *Color Space Transformations* foram aplicados os métodos *Random Brightness*, *Random Contrast*, *Random Saturation* e *Random Lighting*. Já na categoria de *Geometric Transformations* foram aplicados os métodos *Resize Shortest Edge* e *Random Flip*. A Tabela 20 apresenta de forma sumarizada os métodos de aumento de dados aplicados, bem como os parâmetros utilizados em cada método.

Tabela 20 – Parâmetros dos métodos de aumento de dados utilizados nas imagens de fundo.

Método	Descrição	Parâmetro
Random Brightness	Altera aleatoriamente o brilho da imagem	(0,8, 1,2)
Random Contrast	Altera aleatoriamente o contraste da imagem	(0,8, 1,2)
Random Saturation	Altera aleatoriamente a saturação da imagem	(0,8, 1,2)
Random Lighting	Altera aleatoriamente a iluminação da imagem de acordo com o parâmetro de escala	0,7
Resize Shortest Edge	Dimensiona a borda mais curta para o tamanho fornecido com um limite de max_size na borda mais longa.	$short\_edge\_length =$ (640, 672, 704, 736, 768, 800) $max\_size =$ 1333
Random Flip	Vira aleatoriamente as imagens de acordo com uma probabilidade	prob=0,5

Após a etapa de aumento de dados fez-se o treinamento da abordagem para realizar a segmentação de instância das lesões de fundo. Os detalhes sobre a arquitetura da rede neural profunda que compõe a abordagem são apresentados a seguir.

7.2 Arquitetura da Rede Neural Profunda

A abordagem realiza a detecção e a segmentação em termos de *pixel* das lesões presentes em imagens de fundo. Primeiramente, as propostas de região das imagens são verificadas e classificadas. Em seguida, são geradas as caixas delimitadoras e as máscaras de segmentação das lesões identificadas. O processo de criação da

máscara de cada lesão é realizado usando uma rede neural convolucional adicional sobre um mapa de características, em que uma matriz é gerada e preenchida com 1 em todos os locais onde o *pixel* pertence à lesão e 0 como saída nos demais locais.

A arquitetura da rede neural é constituída basicamente por três módulos principais, sendo um *Backbone*, um RPN e um *Box Head*, conforme ilustrado no diagrama de blocos apresentado na Figura 55. O *Backbone* é uma rede neural convolucional convencional e tem por finalidade extrair as características das lesões nas imagens de fundo. Na Detectron2, a resolução da imagem de entrada não necessita ser do mesmo tamanho da entrada do modelo pré-treinado e, portanto, é possível usar *a priori* qualquer resolução de imagem para entrada do *Backbone*.

Entretanto, em função de limitações associadas ao equipamento de hardware utilizado para realizar os experimentos, optou-se por utilizar a resolução padrão da Detectron2, que redimensiona as imagens de entrada de acordo com os parâmetros `INPUT.MIN_SIZE_TRAIN` igual a 800 *pixels* e `INPUT.MAX_SIZE_TRAIN` igual a 1.333 *pixels*. Como as imagens do conjunto de dados DDR possuem largura superior ao tamanho máximo definido, a Detectron2 redimensiona as imagens de fundo para a largura de 1.333 *pixels*, ajustando a altura da imagem proporcionalmente à largura e de acordo o tamanho mínimo definido de 800 *pixels*. É importante destacar que as imagens do conjunto DDR possuem tamanhos (altura e largura) variáveis.

O *Backbone* da arquitetura é composto por uma rede residual com camadas convolucionais agrupadas designada por ResNeXt (XIE et al., 2017). Nos experimentos realizados, a arquitetura ResNeXt-101-32$\times$8d-FPN obteve a melhor precisão na detecção e segmentação das lesões. Esta arquitetura consiste em uma estrutura que contêm múltiplos blocos *bottleneck*. A vantagem da ResNeXt em relação a arquiteturas ResNet tradicionais é a inclusão de uma nova dimensão de cardinalidade, além das dimensões de profundidade e largura. A arquitetura ResNeXt-101-32$\times$8d-FPN é composta por 101 camadas e agrupamento (cardinalidade) de 32 grupos de convoluções com largura de grupo de 8 dimensões ($\sim$ 88 milhões de parâmetros). A cardinalidade implementada na ResNeXt-101-32$\times$8d permite melhorar classificação de imagens sem, no entanto, aumentar o custo computacional associado ao acréscimo de parâmetros à arquitetura da rede neural (XIE et al., 2017). A estrutura do *Backbone* é seguida por uma FPN (LI et al., 2019).

A FPN é utilizada como *Neck* da arquitetura, sendo uma extensão da rede neural profunda utilizada no *Backbone*. O objetivo do *Neck* é realizar a extração de características das lesões em diferentes escalas. A FPN utilizada possui cinco escalas com saídas denominadas P2, P3, P4, P5 e P6, respectivamente, com tamanho de canal C=256, para todas as escalas, e tamanho de *stride* $S = (4, 8, 16, 32, 64)$, respectivamente.

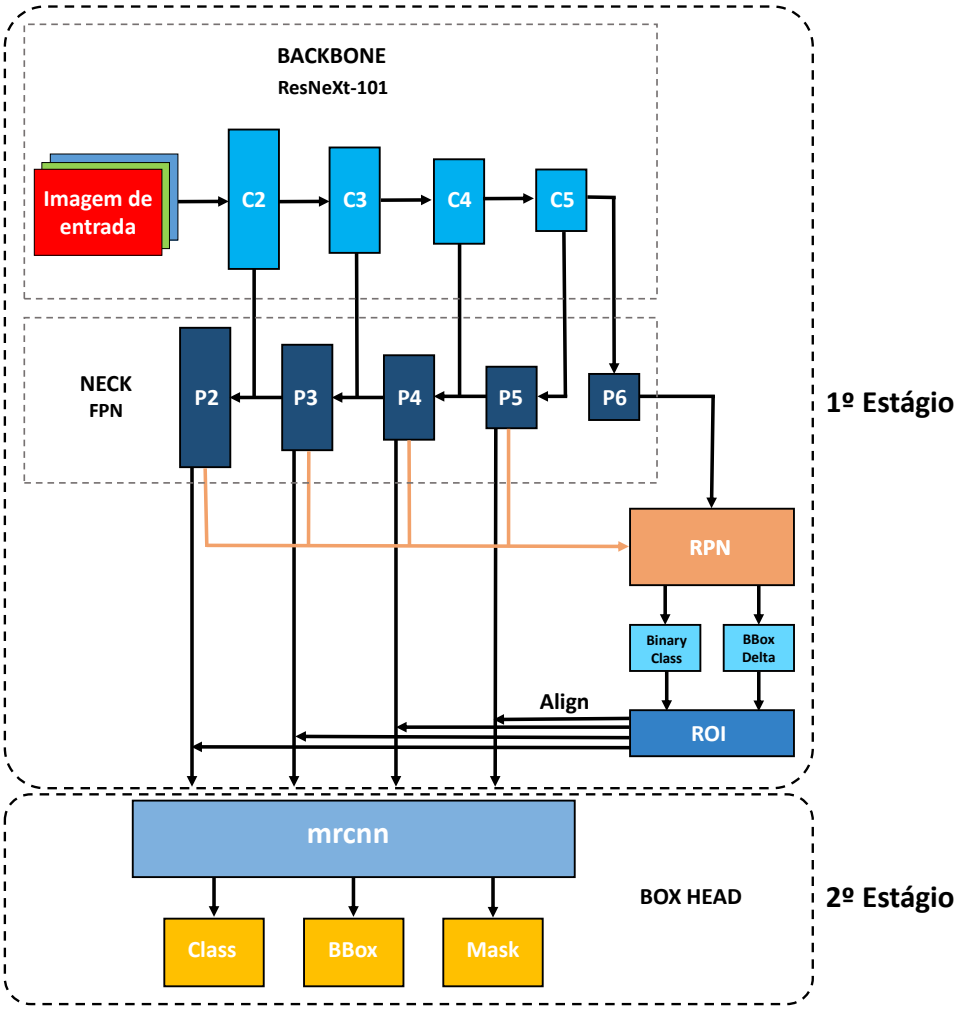


Figura 55 – Diagrama de blocos da arquitetura da Mask R-CNN com dois estágios. No primeiro estágio a arquitetura possui um módulo de *Backbone* composto por uma rede neural profunda ResNeXt-101 e um módulo de *Neck* composto por uma *Feature Pyramid Network* (FPN). Os mapas de características das imagens de fundo são extraídas pelo *Backbone* e encaminhadas para as camadas da FPN que são integradas com um módulo RPN. O segundo estágio da arquitetura é composto por um módulo Box Head, que recebe as regiões selecionadas pelo classificador ROI e gera as *bounding boxes* (BBox) e as máscaras de segmentação (*Mask*) para estas regiões.

Portanto, se apenas uma imagem de entrada de tamanho 1.333×1.333 *pixels* for utilizada na entrada do *Backbone*, os tamanhos dos mapas de características de saída das camadas P2, P3, P4, P5 e P6 serão de $334 \times 334 \times 256$, $166 \times 166 \times 256$, $84 \times 84 \times 256$, $40 \times 40 \times 256$ e $20 \times 20 \times 256$, respectivamente. Desse modo, as camadas P2 e P3 são utilizadas para a detecção de objetos de tamanhos pequenos, ao passo que as camadas P5 e P6 são responsáveis pela detecção de objetos com tamanhos maiores. A FPN extrai mapas de características em várias escalas e com diferentes campos receptivos.

Na sequência, a arquitetura possui um módulo RPN cuja tarefa é fazer a inspeção de cima para baixo em todo o FPN do *Backbone*, a fim de propor regiões que podem conter lesões na imagem de fundo. Na Detectron2 toda a computação realizada pelo RPN é realizada na GPU. O módulo RPN usa âncoras que são um conjunto de caixas com localizações predefinidas dos objetos, dimensionadas de acordo com as imagens de entrada. Âncoras individuais são atribuídas às classes e caixas delimitadoras. O RPN gera duas saídas para cada âncora: a classe da âncora e as especificações da caixa delimitadora. Convém destacar que o RPN detecta regiões a partir de características multi-escala. Por padrão, são obtidas aproximadamente 1.000 propostas de caixa com as pontuações de confiança.

O valor de classificação (Binary Class) e o valor de regressão de caixas delimitadoras (BBox Delta) (vide Figura 55) é gerado por meio do RPN. Os valores dos deltas t são calculados conforme as Equações:

$$t_x = \frac{x - x_a}{w_a} \tag{39}$$

$$t_y = \frac{y - y_a}{h_a} \tag{40}$$

$$t_w = \log \frac{w}{w_a} \tag{41}$$

$$t_h = \log \frac{h}{h_a} \tag{42}$$

onde: t_x e t_y correspondem ao centro das coordenadas; t_w e t_h correspondem à largura e altura, respectivamente; x, y, w e h correspondem às caixas preditas; e, x_a, y_a, w_a e h_a correspondem às caixas de âncora.

O último módulo da arquitetura é o *Box Head*, responsável pelo corte e interpolação dos mapas de características das propostas de região geradas no RPN. Somente as características das camadas P2, P3, P4 e P5 do FPN são utilizadas no *Box Head*. Além disso, o *Box Head* obtém a localização das caixas ajustadas e os resultados da classificação por meio de camadas totalmente conectadas. Este módulo possui um ROI *Pooling*. Segundo He et al. (2020), as regiões do mapa de características selecionadas ficam desalinhadas em relação às propostas de regiões da imagem original. Como a segmentação da imagem requer especificidade em termos de *pixel*, este problema pode ocasionar imprecisões durante a segmentação. Para resolver este problema utilizou-se um ROI *Align* (HE et al., 2020), de forma que o mapa de características seja amostrado em pontos diferentes para então ser aplicada uma interpolação bilinear com o propósito de obter a posição precisa do *pixel*.

Utilizou-se o ROI *Align* V2 (versão modificada do ROI *Align*) que realiza o deslocamento de meio *pixel* (0,5), subtraído das coordenadas de ROI para calcular os índices de *pixels* vizinhos com mais precisão. Após passar pelo ROI *Pooling*, uma rede convolucional é usada para receber as regiões selecionadas pelo classificador ROI e gerar as *bounding boxes* e *masks* para estas regiões. Por padrão, o número de ROIs para um lote n é 512 e o tamanho do ROI é 7$\times$7.

Após o ROI *Pooling*, as características recortadas são utilizadas na arquitetura do *Head*. No caso da Mask R-CNN existem dois tipos de cabeças: *Box Head* e *Mask Head*. O cálculo da função de perda das saídas durante o treinamento é realizado por meio de duas funções: *localization loss* (`loss_box_reg`), obtida por meio da função *Smooth L1 loss*; e, *classification loss* (`loss_cls`), obtida por meio da função de perda de entropia cruzada *Softmax* (WU et al., 2019). Os resultados destas *losses* são adicionados às perdas calculadas no RPN (`loss_rpn_loc` e `loss_rpn_cls`) e somados à `Total Loss` (WU et al., 2019).

Dessa forma, a função de perda total é calculada de acordo com a soma ponderada de diferentes tipos de perdas do modelo atribuídas a cada uma destas etapas: classificação, detecção e segmentação. A seguir, é realizada uma breve explicação das diferentes funções de perda que compõem a perda total da abordagem para segmentação de instância:

`RPN Class Loss`**:** é a perda atribuída à classificação indevida das âncoras (presença ou ausência de qualquer lesão) pelo módulo RPN. O peso desta função deve ser aumentado quando várias lesões não estão sendo detectadas pelo modelo na camada de saída da rede neural. O aumento do peso desta função pode garantir que o módulo RPN capture determinadas lesões que porventura não tenham sido detectadas inicialmente. Mede a perda de *objectness*, ou seja, quão bom o RPN é em rotular as caixas de âncora como primeiro plano ou plano de fundo (REN et al., 2017).

`RPN BBox Loss`**:** é a perda relacionada à precisão de localização do RPN. O ajuste deste peso deve ser realizado caso uma lesão seja detectada, porém a sua caixa delimitadora necessita ser corrigida em função de não estar ajustada adequadamente aos objetos detectados (REN et al., 2017).

`mrcnn Class Loss`**:** é a perda atribuída à classificação inadequada da lesão que está presente na proposta de região. O peso deve ser aumentado caso a lesão esteja sendo detectada na imagem, mas classificada incorretamente. Está associada à perda de classificação no ROI *Head*. Mede a perda por classificação de caixa, ou seja, quão bom o modelo é em rotular uma caixa prevista com a classe correta (GIRSHICK, 2015).

`mrcnn BBox Loss`**:** é a perda atribuída à localização da caixa delimitadora da classe identificada. O peso deve ser aumentado se a classificação da lesão estiver correta, mas a sua localização imprecisa. Está associada à perda de localização no ROI *Head*. Mede a perda de localização da caixa (localização prevista *versus* localização real) (GIRSHICK, 2015).

`mrcnn Mask Loss`**:** é a perda associada às máscaras binárias criadas das lesões identificadas. Se a identificação em termos de *pixel* for importante, este peso deve ser aumentado sempre que possível (HE et al., 2020).

`Total Loss`**:** é a perda associada à soma ponderada das demais perdas explicadas anteriormente, calculadas durante as iterações do modelo durante a etapa de treinamento.

Como mencionado, o modelo Mask R-CNN possui dois estágios. No primeiro estágio regiões de interesse (ROI) são geradas por um módulo RPN e no segundo estágio geram-se as saídas de classe, regressão de caixas delimitadoras e máscaras binárias usando a ROI gerada no primeiro estágio. Para cada região de interesse analisada uma função de perda multitarefa é aplicada (ZIMMERMANN; SIEMS, 2019), conforme a Equação 43:

$$Loss = L_{Class} + L_{BBox} + L_{Mask} \tag{43}$$

onde: L_{Class} corresponde à perda de classificação, L_{BBox} corresponde à perda de regressão da caixa delimitadora e L_{Mask} corresponde à perda da máscara. O resultado da função de perda final é obtido com base na média da perda sobre as amostras.

Além disso, para realizar a inferência das lesões de fundo, uma etapa de pós-processamento é realizada com o objetivo de filtrar as caixas delimitadoras de baixa pontuação. Para tanto, a técnica de supressão de não-máximos é aplicada para eliminar ROIs que estejam abaixo de um limite de pontuação pré-definido.

7.2.1 Pré-treinamento

Nesta etapa foi utilizada transferência de aprendizado para realizar o pré-treinamento da arquitetura de rede neural. Os pesos pré-treinados no conjunto de dados COCO (LIN et al., 2014) foram importados para inicializar os pesos da arquitetura de rede neural utilizada na abordagem. Após realizar o pré-treinamento, a saída da rede neural foi modificada para se adequar à segmentação de instância das lesões retinianas associadas à RD, preservando os pesos das camadas superiores.

Com o pré-treinamento foi possível reduzir o custo computacional da abordagem durante a etapa de treinamento. A realização deste processo baseou-se na metodologia proposta por Franke et al. (2021) em que: (1) as camadas iniciais da arquitetura são pré-treinadas com os pesos do conjunto de dados COCO; (2) as últimas camadas são cortadas e substituídas por novas camadas; (3) as camadas adicionadas são ajustadas no conjunto de dados de RD; e, (4) toda a rede neural é treinada novamente após o ajuste fino das camadas finais da arquitetura, para que sejam realizados pequenos ajustes nos pesos de toda a arquitetura.

7.2.2 Treinamento e Ajuste do Modelo

Na abordagem para segmentação de instância foram utilizadas as imagens dos conjuntos de dados públicos DDR e IDRiD. Estes conjuntos de dados foram divididos, para os experimentos, em conjuntos de treino, validação e teste em uma proporção de 50:20:30, respectivamente. O conjunto de treino foi utilizado para treinar o modelo, o conjunto de validação foi utilizado para realizar a comparação de diferentes modelos e hiperparâmetros e o conjunto de teste foi utilizado para avaliar se o modelo ajustado na etapa de validação generaliza adequadamente mesmo sobre dados não conhecidos *a priori*, objetivando garantir a aleatoriedade do conjunto de dados, reduzindo vieses nos resultados produzidos por modelos avaliados sobre dados já utilizados em etapas de treinamento e ajuste de hiperparâmetros.

A etapa de ajuste fino teve como objetivo otimizar os hiperparâmetros buscando resultados mais precisos na segmentação de instância das lesões de fundo. A metodologia adotada para realizar o ajuste fino de hiperparâmetros consistiu das seguintes etapas: (1) para cada ajuste realizado, varia-se um valor de hiperparâmetro e a abordagem é retreinada, mantendo constantes os demais valores de hiperparâmetros; (2) avalia-se o efeito desta otimização por meio da avaliação do desempenho com as métricas *Average Precision* (AP) e *mean Average Precision* (mAP); (3) ajusta-se (aumento ou diminuindo) o valor do hiperparâmetro caso uma melhoria nos valores das métricas seja obtido até que o máximo local seja alcançado; e, (4) o mesmo processo é realizado para os demais hiperparâmetros até que seja obtido um conjunto ótimo de hiperparâmetros que produzam resultados máximos de AP e mAP.

Após a realização destas etapas no conjunto de validação do *dataset* DDR foi encontrado o melhor ajuste de hiperparâmetros, conforme apresentado na Tabela 21. A próxima etapa foi avaliar a abordagem no conjunto de teste dos *datasets* DDR e IDRiD, a fim de verificar a capacidade de generalização da abordagem em imagens de fundo não conhecidas *a priori*.

Tabela 21 – Hiperparâmetros ajustados da abordagem para segmentação de instância na etapa de validação utilizando o conjunto de dados DDR.

Hiperparâmetro	**Valor**
Early Stopping	*Patience value* = 100
Images per Batch	2
Max Iterations	50.000
Momentum	0,937
NMS Testing Threshold	0,5
Number of Workers	4
Otimizador	SGD e Adam
ROI Heads Batch Size per Image	1.024
RPN Batch Size per Image	512
RPN Positive Fraction	0,5
Tamanho das âncoras ajustadas	(8, 16, 32, 64, 128)
Taxa de Aprendizado	0,001 - 0,0001
Test Detections per Image	256
Weight Decay	0,0005

O melhor ajuste obtido para a abordagem inclui *Number of Workers* igual a 4, *Images per Batch* igual a 2, *Max iterations* igual a 50.000, e Taxa de Aprendizado igual a 0,0001. O tamanho das âncoras de células foi definido como (8, 16, 32, 64, 128). Para reduzir o subajuste do modelo, na etapa de treinamento, a abordagem foi regularizada por meio da criação de imagens artificiais, a fim de aumentar a quantidade de exemplos das lesões anotadas. Com o intuito de minimizar a probabilidade de sobreajuste da rede neural, principalmente em função do desbalanceamento da quantidade de exemplos das diferentes classes de lesões, foi aplicada a técnica de regularização L2 (GOODFELLOW et al., 2020) por meio da introdução do termo de penalidade *Weight Decay* durante a etapa de treinamento do modelo.

A seguir são explicados os principais hiperparâmetros ajustados na abordagem para a realização da segmentação de instância de lesões retinianas associadas à RD, conforme apresentado na Tabela 21:

***Images per Batch*:** se refere ao número de imagens de treinamento por lote.

***Max Iterations*:** se refere ao número máximo de iterações realizadas durante o treinamento da rede. A iteração é o número de lotes ou etapas de dados de treinamento necessários para completar uma época.

***NMS Testing Threshold*:** se refere ao limite de Interseção sobre União usado para supressão não máxima de caixas delimitadoras.

***Number of Workers*:** se refere ao número de *threads* de carregamento de dados.

***ROI Heads Batch Size per Image*:** se refere ao número máximo de ROIs que o módulo RPN gera para a imagem a fim de ser processada para a classificação e segmentação. A maneira mais adequada para definir o valor deste hiperparâmetro é iniciar com o valor padrão se o número de instâncias na imagem for desconhecido. Caso o número de instâncias seja limitado o valor para este hiperparâmetro pode ser diminuído para reduzir o tempo de treinamento.

***ROI Heads Positive Fraction*:** se refere ao limite do nível de confiança, ou seja, a partir deste valor ocorrerá a classificação de uma instância. Geralmente, a inicialização do percentual de confiança pode ser padrão e reduzida ou aumentada com base no número de instâncias detectadas. Caso a detecção de todos os objetos seja importante e os falsos positivos forem adequados, o limite pode ser reduzido para identificar todas as instâncias possíveis. No entanto, caso a precisão da detecção seja importante é necessário aumentar o limite para garantir que haja um mínimo de falsos positivos de forma que o modelo faça a predição apenas das instâncias com um nível de confiança mais elevado.

***RPN Batch Size per Image*:** se refere ao número de regiões por imagem usadas para treinar o módulo RPN.

Tamanho das Âncoras: se refere aos tamanhos de âncora em *pixels* absolutos para a entrada da rede.

***Test Detections per Image*:** se refere ao número máximo de instâncias que podem ser detectadas em uma imagem. A configuração deste parâmetro influencia na redução de falsos positivos, além de reduzir o tempo de treinamento.

7.3 Experimentos, Resultados e Discussões

A abordagem possui uma arquitetura Mask R-CNN com um *Backbone* ResNeXt-101-32$\times$8d-FPN, construída por meio da biblioteca de código aberto Detectron2 e pré-treinada no conjunto de dados COCO. Para realizar os experimentos os conjuntos de dados DDR e IDRiD foram divididos em conjunto de dado de treino, validação e teste em uma proporção de 50:20:30, respectivamente.

O conjunto de validação foi utilizado para realizar o ajuste-fino dos hiperparâmetros da arquitetura e o conjunto de teste foi utilizado para realizar a avaliação da abordagem em imagens não conhecidas *a priori*. A avaliação foi realizada tanto na detecção quanto na segmentação das lesões de fundo. Os resultados obtidos pela abordagem foram comparados com trabalhos de mesmo propósito encontrados na literatura. Para avaliar a qualidade das detecções (BBox) e segmentações (Mask) adotou-se o limite de IoU de 0,5 durante os experimentos. Também foram utilizadas as métricas AP e

mAP para avaliar a precisão da abordagem na segmentação de instância das lesões de fundo.

A abordagem para segmentação de instância foi comparada com diferentes modelos que utilizam redes neurais densas de última geração. Nos experimentos foram utilizados os modelos: 1) Mask R-CNN ResNet-50 C4 (HE et al., 2020), que usa um *Backbone* ResNet conv4 com cabeça conv5 (REN et al., 2017); 2) Mask R-CNN ResNet-50 C5-dilated (HE et al., 2020), que usa *Backbone* ResNet conv5, com dilatações em conv5 e cabeças padrão conv e FC para previsão de máscara e caixa delimitadora, respectivamente (DAI et al., 2017); 3) Mask R-CNN ResNet-50 FPN (HE et al., 2020), que usa um *Backbone* ResNet$\times$FPN com cabeças padrão conv e FC para a previsão de máscara e caixa delimitadora, respectivamente; 4) Mask R-CNN ResNet-101 C4 (HE et al., 2020); 5) Mask R-CNN ResNet-101 C5-dilated (HE et al., 2020); 6) Mask R-CNN ResNet-101 FPN (HE et al., 2020); e, 7) Mask R-CNN ResNeXt-101-32$\times$8d-FPN, conforme apresentado nas Tabelas 22, 23, 24, 25, 26, 27, 28 e 29. A seguir, são apresentados e discutidos os resultados obtidos.

A Tabela 22 apresenta os resultados obtidos com a métrica mAP para o limite de IoU de 0,5 com otimizador SGD. A abordagem utilizando o *Backbone* ResNeXt-101-32$\times$8d-FPN e *Tilling* alcançou a melhor precisão nas tarefas de detecção e segmentação nos experimentos realizados no conjunto de validação do *dataset* DDR, conforme destacado em negrito, com um mAP de 0,2660 na detecção das caixas delimitadoras (BBox) das lesões e mAP de 0,2600 na identificação das máscaras de segmentação (Mask) das lesões de fundo.

Tabela 22 – Resultados obtidos nas tarefas de detecção e segmentação das lesões de fundo com a métrica mAP para o limite de IoU de 0,5 no conjunto de validação do *dataset* DDR com otimizador SGD.

	Modelos	**Backbone**	*mAP*
BBox	Mask R-CNN (*baseline*) (WU et al., 2019)	ResNet-50 C4	0,1766
	Mask R-CNN (*baseline*) (WU et al., 2019)	ResNet-50 C5-dilated	0,1368
	Mask R-CNN (*baseline*) (WU et al., 2019)	ResNet-50 FPN	0,1737
	Mask R-CNN (*baseline*) (WU et al., 2019)	ResNet-101 C4	0,1746
	Mask R-CNN (*baseline*) (WU et al., 2019)	ResNet-101 C5-dilated	0,1678
	Mask R-CNN (*baseline*) (WU et al., 2019)	ResNet-101 FPN	0,1909
	Mask R-CNN (*baseline*) (WU et al., 2019)	ResNeXt-101-32$\times$8d-FPN	0,2125
	Abordagem para Segmentação de Instância sem *Tilling*	ResNeXt-101-32$\times$8d-FPN	0,1639
	Abordagem para Segmentação de Instância com *Tilling*	**ResNeXt-101-32$\times$8d-FPN**	**0,2660**
Mask	Mask R-CNN (*baseline*) (WU et al., 2019)	ResNet-50 C4	0,1617
	Mask R-CNN (*baseline*) (WU et al., 2019)	ResNet-50 C5-dilated	0,1328
	Mask R-CNN (*baseline*) (WU et al., 2019)	ResNet-50 FPN	0,1795
	Mask R-CNN (*baseline*) (WU et al., 2019)	ResNet-101 C4	0,1625
	Mask R-CNN (*baseline*) (WU et al., 2019)	ResNet-101 C5-dilated	0,1603
	Mask R-CNN (*baseline*) (WU et al., 2019)	ResNet-101 FPN	0,1848
	Mask R-CNN (*baseline*) (WU et al., 2019)	ResNeXt-101-32$\times$8d-FPN	0,2206
	Abordagem para Segmentação de Instância sem *Tilling*	ResNeXt-101-32$\times$8d-FPN	0,1651
	Abordagem para Segmentação de Instância com *Tilling*	**ResNeXt-101-32$\times$8d-FPN**	**0,2600**

Já a Tabela 23 apresenta os resultados obtidos com a métrica mAP para o limite de IoU de 0,5 com otimizador SGD no conjunto de teste do *dataset* DDR. A abordagem com o *Backbone* ResNeXt-101-32×8d-FPN e *Tilling* obteve a melhor precisão nas tarefas de detecção e segmentação neste experimento, conforme destacado em negrito, com um mAP de 0,1600 na detecção e 0,1630 na segmentação das máscaras das lesões de fundo.

Tabela 23 – Resultados obtidos nas tarefas de detecção e segmentação das lesões de fundo com a métrica mAP para o limite de IoU de 0,5 no conjunto de teste do *dataset* DDR com otimizador SGD.

	Modelos	**Backbone**	*mAP*
BBox	Mask R-CNN (*baseline*) (WU et al., 2019)	ResNet-50 C4	0,0940
	Mask R-CNN (*baseline*) (WU et al., 2019)	ResNet-50 C5-dilated	0,0681
	Mask R-CNN (*baseline*) (WU et al., 2019)	ResNet-50 FPN	0,0964
	Mask R-CNN (*baseline*) (WU et al., 2019)	ResNet-101 C4	0,0821
	Mask R-CNN (*baseline*) (WU et al., 2019)	ResNet-101 C5-dilated	0,0821
	Mask R-CNN (*baseline*) (WU et al., 2019)	ResNet-101 FPN	0,0892
	Mask R-CNN (*baseline*) (WU et al., 2019)	ResNeXt-101-32×8d-FPN	0,1153
	Abordagem para Segmentação de Instância sem *Tilling*	ResNeXt-101-32×8d-FPN	0,0860
	Abordagem para Segmentação de Instância com *Tilling*	**ResNeXt-101-32×8d-FPN**	**0,1600**
Mask	Mask R-CNN (*baseline*) (WU et al., 2019)	ResNet-50 C4	0,0865
	Mask R-CNN (*baseline*) (WU et al., 2019)	ResNet-50 C5-dilated	0,0636
	Mask R-CNN (*baseline*) (WU et al., 2019)	ResNet-50 FPN	0,1025
	Mask R-CNN (*baseline*) (WU et al., 2019)	ResNet-101 C4	0,0769
	Mask R-CNN (*baseline*) (WU et al., 2019)	ResNet-101 C5-dilated	0,0737
	Mask R-CNN (*baseline*) (WU et al., 2019)	ResNet-101 FPN	0,0857
	Mask R-CNN (*baseline*) (WU et al., 2019)	ResNeXt-101-32×8d-FPN	0,1172
	Abordagem para Segmentação de Instância sem *Tilling*	ResNeXt-101-32×8d-FPN	0,0884
	Abordagem para Segmentação de Instância com *Tilling*	**ResNeXt-101-32×8d-FPN**	**0,1630**

A Tabela 24 apresenta os resultados obtidos pela abordagem com a métrica mAP para o limite de IoU de 0,5 com otimizador Adam no conjunto de validação do dataset DDR. Mais uma vez o melhor resultado foi obtido com o *Backbone* ResNeXt-101-32×8d-FPN e *Tilling*, conforme destacado em negrito, com mAP de 0,2903 na detecção e mAP de 0,2941 na segmentação das lesões.

Já a Tabela 25 apresenta os resultados obtidos com a métrica mAP para o limite de IoU de 0,5 com otimizador Adam no conjunto de teste do *dataset* DDR. Com o *Backbone* ResNeXt-101-32×8d-FPN a abordagem alcançou os melhores resultados na detecção e segmentação das lesões, com mAP de 0,1670 e 0,1735, respectivamente, conforme destacado em negrito.

Também foram avaliados e comparados os resultados obtidos pela abordagem no conjunto de dados IDRiD, conforme apresentado nas Tabelas 26, 27, 28 e 29. Como nos experimentos realizados com a abordagem no *dataset* DDR, o *Backbone* ResNeXt-101-32×8d-FPN obteve os melhores resultados nas tarefas de detecção e segmentação das lesões de fundo. Entretanto, é importante destacar que no conjunto de dados IDRiD os melhores resultados obtidos pela abordagem proposta com o otimi-

Tabela 24 – Resultados obtidos nas tarefas de detecção e segmentação das lesões de fundo com a métrica mAP para o limite de IoU de 0,5 no conjunto de validação do *dataset* DDR com otimizador Adam.

	Modelos	**Backbone**	*mAP*
BBox	Mask R-CNN (*baseline*) (WU et al., 2019)	ResNet-50 C4	0,1982
	Mask R-CNN (*baseline*) (WU et al., 2019)	ResNet-50 C5-dilated	0,1833
	Mask R-CNN (*baseline*) (WU et al., 2019)	ResNet-50 FPN	0,1917
	Mask R-CNN (*baseline*) (WU et al., 2019)	ResNet-101 C4	0,1821
	Mask R-CNN (*baseline*) (WU et al., 2019)	ResNet-101 C5-dilated	0,1775
	Mask R-CNN (*baseline*) (WU et al., 2019)	ResNet-101 FPN	0,1975
	Mask R-CNN (*baseline*) (WU et al., 2019)	ResNeXt-101-32×8d-FPN	0,2078
	Abordagem para Segmentação de Instância sem *Tilling*	ResNeXt-101-32×8d-FPN	0,2354
	Abordagem para Segmentação de Instância com *Tilling*	**ResNeXt-101-32×8d-FPN**	**0,2903**
Mask	Mask R-CNN (*baseline*) (WU et al., 2019)	ResNet-50 C4	0,1901
	Mask R-CNN (*baseline*) (WU et al., 2019)	ResNet-50 C5-dilated	0,1641
	Mask R-CNN (*baseline*) (WU et al., 2019)	ResNet-50 FPN	0,1983
	Mask R-CNN (*baseline*) (WU et al., 2019)	ResNet-101 C4	0,1616
	Mask R-CNN (*baseline*) (WU et al., 2019)	ResNet-101 C5-dilated	0,1612
	Mask R-CNN (*baseline*) (WU et al., 2019)	ResNet-101 FPN	0,1965
	Mask R-CNN (*baseline*) (WU et al., 2019)	ResNeXt-101-32×8d-FPN	0,2144
	Abordagem para Segmentação de Instância sem *Tilling*	ResNeXt-101-32×8d-FPN	0,2350
	Abordagem para Segmentação de Instância com *Tilling*	**ResNeXt-101-32×8d-FPN**	**0,2941**

Tabela 25 – Resultados obtidos nas tarefas de detecção e segmentação das lesões de fundo com a métrica mAP para o limite de IoU de 0,5 no conjunto de teste do *dataset* DDR com otimizador Adam.

	Modelos	**Backbone**	*mAP*
BBox	Mask R-CNN (*baseline*) (WU et al., 2019)	ResNet-50 C4	0,1055
	Mask R-CNN (*baseline*) (WU et al., 2019)	ResNet-50 C5-dilated	0,0953
	Mask R-CNN (*baseline*) (WU et al., 2019)	ResNet-50 FPN	0,1071
	Mask R-CNN (*baseline*) (WU et al., 2019)	ResNet-101 C4	0,0973
	Mask R-CNN (*baseline*) (WU et al., 2019)	ResNet-101 C5-dilated	0,0961
	Mask R-CNN (*baseline*) (WU et al., 2019)	ResNet-101 FPN	0,1127
	Mask R-CNN (*baseline*) (WU et al., 2019)	ResNeXt-101-32×8d-FPN	0,1278
	Abordagem para Segmentação de Instância sem *Tilling*	ResNeXt-101-32×8d-FPN	0,1390
	Abordagem para Segmentação de Instância com *Tilling*	**ResNeXt-101-32×8d-FPN**	**0,1670**
Mask	Mask R-CNN (*baseline*) (WU et al., 2019)	ResNet-50 C4	0,1011
	Mask R-CNN (*baseline*) (WU et al., 2019)	ResNet-50 C5-dilated	0,0797
	Mask R-CNN (*baseline*) (WU et al., 2019)	ResNet-50 FPN	0,1093
	Mask R-CNN (*baseline*) (WU et al., 2019)	ResNet-101 C4	0,0823
	Mask R-CNN (*baseline*) (WU et al., 2019)	ResNet-101 C5-dilated	0,0847
	Mask R-CNN (*baseline*) (WU et al., 2019)	ResNet-101 FPN	0,1071
	Mask R-CNN (*baseline*) (WU et al., 2019)	ResNeXt-101-32×8d-FPN	0,1295
	Abordagem para Segmentação de Instância sem *Tilling*	ResNeXt-101-32×8d-FPN	0,1349
	Abordagem para Segmentação de Instância com *Tilling*	**ResNeXt-101-32×8d-FPN**	**0,1735**

zador SGD na etapa de validação foram com a utilização de *Tilling*, alcançando mAP de 0,3460 na detecção e 0,3210 na segmentação (indicado em negrito), ao passo que os melhores resultados obtidos pela abordagem proposta com o otimizador Adam na etapa de validação foram sem a utilização de *Tilling*, com mAP de 0,3127 na detecção e 0,3041 na segmentação das lesões de fundo, conforme indicado em negrito.

Os resultados obtidos pela abordagem no conjunto de dados IDRiD foram superiores aos obtidos no DDR. Uma possível explicação para estes resultados está relacionada à qualidade e resolução das imagens disponíveis nestes *datasets*, uma vez que no IDRiD as imagens possuem resolução de 4288×2848 *pixels*, ao passo que no *dataset* DDR as imagens possuem resoluções com tamanhos variáveis, o que acaba por impactar na extração de características das lesões presentes nestas imagens. Os resultados também indicam que com a aplicação de *Tilling* houve um aumento na precisão da detecção das lesões em ambos *datasets*, apresentando somente a exceção quando aplicado no conjunto de dados IDRiD com a utilização de otimizador Adam.

Tabela 26 – Resultados obtidos nas tarefas de detecção e segmentação das lesões de fundo com a métrica mAP para o limite de IoU de 0,5 no conjunto de validação do *dataset* IDRiD com otimizador SGD.

	Modelos	**Backbone**	*mAP*
BBox	Mask R-CNN (*baseline*) (WU et al., 2019)	ResNet-50 C4	0,2254
	Mask R-CNN (*baseline*) (WU et al., 2019)	ResNet-50 C5-dilated	0,1789
	Mask R-CNN (*baseline*) (WU et al., 2019)	ResNet-50 FPN	0,2050
	Mask R-CNN (*baseline*) (WU et al., 2019)	ResNet-101 C4	0,2120
	Mask R-CNN (*baseline*) (WU et al., 2019)	ResNet-101 C5-dilated	0,2065
	Mask R-CNN (*baseline*) (WU et al., 2019)	ResNet-101 FPN	0,1656
	Mask R-CNN (*baseline*) (WU et al., 2019)	ResNeXt-101-32×8d-FPN	0,2188
	Abordagem para Segmentação de Instância sem *Tilling*	ResNeXt-101-32×8d-FPN	0,2955
	Abordagem para Segmentação de Instância com *Tilling*	**ResNeXt-101-32×8d-FPN**	**0,3460**
Mask	Mask R-CNN (*baseline*) (WU et al., 2019)	ResNet-50 C4	0,1921
	Mask R-CNN (*baseline*) (WU et al., 2019)	ResNet-50 C5-dilated	0,1645
	Mask R-CNN (*baseline*) (WU et al., 2019)	ResNet-50 FPN	0,2138
	Mask R-CNN (*baseline*) (WU et al., 2019)	ResNet-101 C4	0,2161
	Mask R-CNN (*baseline*) (WU et al., 2019)	ResNet-101 C5-dilated	0,1841
	Mask R-CNN (*baseline*) (WU et al., 2019)	ResNet-101 FPN	0,1789
	Mask R-CNN (*baseline*) (WU et al., 2019)	ResNeXt-101-32×8d-FPN	0,2051
	Abordagem para Segmentação de Instância sem *Tilling*	ResNeXt-101-32×8d-FPN	0,2699
	Abordagem para Segmentação de Instância com *Tilling*	**ResNeXt-101-32×8d-FPN**	**0,3210**

Tabela 27 – Resultados obtidos nas tarefas de detecção e segmentação das lesões de fundo com a métrica mAP para o limite de IoU de 0,5 no conjunto de teste do *dataset* IDRiD com otimizador SGD.

	Modelos	**Backbone**	*mAP*
BBox	Mask R-CNN (*baseline*) (WU et al., 2019)	ResNet-50 C4	0,2061
	Mask R-CNN (*baseline*) (WU et al., 2019)	ResNet-50 C5-dilated	0,1829
	Mask R-CNN (*baseline*) (WU et al., 2019)	ResNet-50 FPN	0,1924
	Mask R-CNN (*baseline*) (WU et al., 2019)	ResNet-101 C4	0,2347
	Mask R-CNN (*baseline*) (WU et al., 2019)	ResNet-101 C5-dilated	0,2519
	Mask R-CNN (*baseline*) (WU et al., 2019)	ResNet-101 FPN	0,1332
	Mask R-CNN (*baseline*) (WU et al., 2019)	ResNeXt-101-32×8d-FPN	0,1912
	Abordagem para Segmentação de Instância sem *Tilling*	ResNeXt-101-32×8d-FPN	0,3064
	Abordagem para Segmentação de Instância com *Tilling*	**ResNeXt-101-32×8d-FPN**	**0,3660**
Mask	Mask R-CNN (*baseline*) (WU et al., 2019)	ResNet-50 C4	0,2026
	Mask R-CNN (*baseline*) (WU et al., 2019)	ResNet-50 C5-dilated	0,1627
	Mask R-CNN (*baseline*) (WU et al., 2019)	ResNet-50 FPN	0,1885
	Mask R-CNN (*baseline*) (WU et al., 2019)	ResNet-101 C4	0,2189
	Mask R-CNN (*baseline*) (WU et al., 2019)	ResNet-101 C5-dilated	0,2336
	Mask R-CNN (*baseline*) (WU et al., 2019)	ResNet-101 FPN	0,1459
	Mask R-CNN (*baseline*) (WU et al., 2019)	ResNeXt-101-32×8d-FPN	0,1861
	Abordagem para Segmentação de Instância sem *Tilling*	ResNeXt-101-32×8d-FPN	0,2877
	Abordagem para Segmentação de Instância com *Tilling*	**ResNeXt-101-32×8d-FPN**	**0,3420**

Tabela 28 – Resultados obtidos nas tarefas de detecção e segmentação das lesões de fundo com a métrica mAP para o limite de IoU de 0,5 no conjunto de validação do *dataset* IDRiD com otimizador Adam.

	Modelos	**Backbone**	*mAP*
BBox	Mask R-CNN (*baseline*) (WU et al., 2019)	ResNet-50 C4	0,2052
	Mask R-CNN (*baseline*) (WU et al., 2019)	ResNet-50 C5-dilated	0,2280
	Mask R-CNN (*baseline*) (WU et al., 2019)	ResNet-50 FPN	0,1957
	Mask R-CNN (*baseline*) (WU et al., 2019)	ResNet-101 C4	0,1833
	Mask R-CNN (*baseline*) (WU et al., 2019)	ResNet-101 C5-dilated	0,1981
	Mask R-CNN (*baseline*) (WU et al., 2019)	ResNet-101 FPN	0,2101
	Mask R-CNN (*baseline*) (WU et al., 2019)	ResNeXt-101-32×8d-FPN	0,2013
	Abordagem para Segmentação de Instância sem *Tilling*	**ResNeXt-101-32×8d-FPN**	**0,3127**
	Abordagem para Segmentação de Instância com *Tilling*	ResNeXt-101-32×8d-FPN	0,3063
Mask	Mask R-CNN (*baseline*) (WU et al., 2019)	ResNet-50 C4	0,1909
	Mask R-CNN (*baseline*) (WU et al., 2019)	ResNet-50 C5-dilated	0,1985
	Mask R-CNN (*baseline*) (WU et al., 2019)	ResNet-50 FPN	0,1900
	Mask R-CNN (*baseline*) (WU et al., 2019)	ResNet-101 C4	0,1671
	Mask R-CNN (*baseline*) (WU et al., 2019)	ResNet-101 C5-dilated	0,1581
	Mask R-CNN (*baseline*) (WU et al., 2019)	ResNet-101 FPN	0,1942
	Mask R-CNN (*baseline*) (WU et al., 2019)	ResNeXt-101-32×8d-FPN	0,2109
	Abordagem para Segmentação de Instância sem *Tilling*	**ResNeXt-101-32×8d-FPN**	**0,3041**
	Abordagem para Segmentação de Instância com *Tilling*	ResNeXt-101-32×8d-FPN	0,2862

Tabela 29 – Resultados obtidos nas tarefas de detecção e segmentação das lesões de fundo com a métrica mAP para o limite de IoU de 0,5 no conjunto de teste do *dataset* IDRiD com otimizador Adam.

	Modelos	**Backbone**	*mAP*
BBox	Mask R-CNN (*baseline*) (WU et al., 2019)	ResNet-50 C4	0,2173
	Mask R-CNN (*baseline*) (WU et al., 2019)	ResNet-50 C5-dilated	0,2349
	Mask R-CNN (*baseline*) (WU et al., 2019)	ResNet-50 FPN	0,2159
	Mask R-CNN (*baseline*) (WU et al., 2019)	ResNet-101 C4	0,2248
	Mask R-CNN (*baseline*) (WU et al., 2019)	ResNet-101 C5-dilated	0,1831
	Mask R-CNN (*baseline*) (WU et al., 2019)	ResNet-101 FPN	0,2126
	Mask R-CNN (*baseline*) (WU et al., 2019)	ResNeXt-101-32$\times$8d-FPN	0,2093
	Abordagem para Segmentação de Instância sem *Tilling*	**ResNeXt-101-32$\times$8d-FPN**	**0,3233**
	Abordagem para Segmentação de Instância com *Tilling*	ResNeXt-101-32$\times$8d-FPN	0,3042
Mask	Mask R-CNN (*baseline*) (WU et al., 2019)	ResNet-50 C4	0,1980
	Mask R-CNN (*baseline*) (WU et al., 2019)	ResNet-50 C5-dilated	0,2116
	Mask R-CNN (*baseline*) (WU et al., 2019)	ResNet-50 FPN	0,2079
	Mask R-CNN (*baseline*) (WU et al., 2019)	ResNet-101 C4	0,2020
	Mask R-CNN (*baseline*) (WU et al., 2019)	ResNet-101 C5-dilated	0,1543
	Mask R-CNN (*baseline*) (WU et al., 2019)	ResNet-101 FPN	0,2083
	Mask R-CNN (*baseline*) (WU et al., 2019)	ResNeXt-101-32$\times$8d-FPN	0,1960
	Abordagem para Segmentação de Instância sem *Tilling*	**ResNeXt-101-32$\times$8d-FPN**	**0,3117**
	Abordagem para Segmentação de Instância com *Tilling*	ResNeXt-101-32$\times$8d-FPN	0,2852

Na Tabela 30 foram comparados os resultados obtidos pela abordagem com e sem a utilização de *Tilling* na detecção e segmentação de cada tipo de lesão utilizando a métrica AP para o limite de IoU de 0,5 no conjunto de validação do *dataset* DDR utilizando otimizador Adam. Na detecção e com a utilização de *Tilling* a abordagem obteve os valores de AP@0,5 iguais a 0,2941, 0,3508, 0,1644 e 0,3520, para Exsudatos Duros (EX), Exsudatos Algodonosos (SE), Microaneurismas (MA) e Hemorragias (HE), respectivamente.

Tabela 30 – Resultados obtidos pela abordagem na detecção (BBox) e Segmentação (Mask) com as métricas AP e mAP limite de IoU de 0,5 no conjunto de validação do *dataset* DDR utilizando otimizador Adam.

Modelo		*AP*				*mAP*
		EX	**SE**	**MA**	**HE**	
Abordagem para Segmentação de Instância sem *Tilling*	BBox	0,2598	0,3147	0,1087	0,2585	0,2354
	Mask	0,2467	0,3400	0,1190	0,2342	0,2350
Abordagem para Segmentação de Instância com *Tilling*	BBox	0,2941	**0,3508**	0,1644	**0,3520**	0,2903
	Mask	**0,3012**	0,3505	**0,2036**	0,3211	**0,2941**

Já na Tabela 31 foram comparados os resultados obtidos pela abordagem com e sem a utilização de *Tilling* na detecção e segmentação das lesões por meio da métrica AP com limite de IoU 0,5 no conjunto de teste do *dataset* DDR utilizando otimizador Adam. Na detecção e com a utilização de *Tilling* foram obtidos os valores de AP@0,5 iguais a 0,2515, 0,1548, 0,1042 e 0,1577, para Exsudatos Duros (EX), Exsudatos Algodonosos (SE), Microaneurismas (MA) e Hemorragias (HE), respectivamente.

Tabela 31 – Resultados obtidos pela abordagem na detecção (BBox) e Segmentação (Mask) com as métricas AP e mAP com limite de IoU de 0,5 no conjunto de teste do *dataset* DDR utilizando otimizador Adam.

Modelo		*AP*				*mAP*
		EX	**SE**	**MA**	**HE**	
Abordagem para Segmentação de Instância sem *Tilling*	BBox	0,2190	0,1589	0,0691	0,1091	0,1390
	Mask	0,2119	**0,1615**	0,0704	0,0959	0,1349
Abordagem para Segmentação de Instância com *Tilling*	BBox	0,2515	0,1548	0,1042	**0,1577**	0,1670
	Mask	**0,2687**	0,1589	**0,1274**	0,1388	**0,1735**

A Figura 56 apresenta os gráficos referentes aos resultados obtidos pela abordagem com a utilização de *Tilling* durante as etapas de treinamento e validação no conjunto de dados DDR com otimizador Adam, conforme os resultados apresentados nas Tabelas 30 e 31.

A Figura 56(a) apresenta a curva de Loss total comparada com a acurácia da detecção das classes de lesões durante o treinamento. Nas Figuras 56(b) e 56(c) são apresentados os resultados obtidos com a detecção de caixas delimitadoras (BBox) e segmentação (Mask) das lesões de fundo para o AP@0,5 na etapa de validação, em que é possível observar que a maior precisão obtida pela abordagem foi alcançada em torno das 30.000 iterações durante o treinamento. Já na Figura 56(d) são

apresentados os resultados obtidos com a detecção de caixas delimitadoras (BBox) por tipo de lesão de fundo para o AP@0,5 na etapa de validação. É possível verificar que houve maior precisão na detecção de Hemorragias e Exsudatos Algodonosos e menor precisão na detecção de Microaneurismas.

Na Figura 56(a) é possível observar que as curvas convergem rapidamente em função da curva em laranja apresentar somente a acurácia da classificação das lesões durante a etapa de treinamento do modelo, não significando, portanto, que houve *overfitting* do modelo neste momento, pois nem a detecção das caixas delimitadoras das lesões ou na segmentação das máscaras é analisado nesta curva. Já nas Figuras 56(b) e 56(c) é possível verificar que houve de fato uma queda repentina de AP por volta das 45.000 iterações, tanto na detecção quando na segmentação das lesões, causado pelo sobreajuste do modelo.

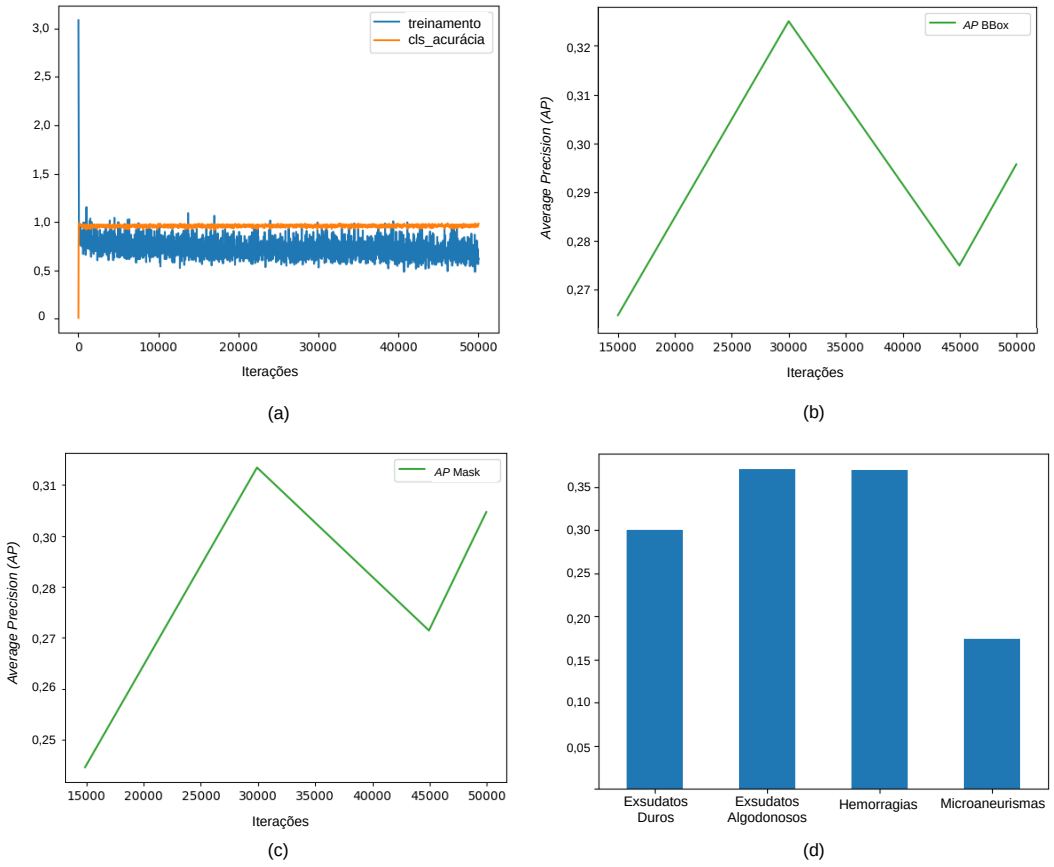


Figura 56 – Treinamento e validação da abordagem para segmentação de instância das lesões de fundo usando o *dataset* DDR com otimizador Adam: (a) curva de *Loss* total *versus* Acurácia da detecção das classes de lesões durante o treinamento; (b) *Average Precision* (AP) da detecção de caixas delimitadoras (BBox) das lesões para o limite de Interseção sobre União (IoU) de 0,5 no conjunto de validação; (c) *Average Precision* (AP) da segmentação de máscaras (Mask) das lesões para o limite de Interseção sobre União (IoU) de 0,5 no conjunto de validação; e, (d) *Average Precision* (AP) da detecção de caixas delimitadoras (BBox) das lesões para o limite de Interseção sobre União (IoU) de 0,5:0,95 no conjunto de validação.

O tempo médio de inferência para realizar a segmentação de instância das lesões de fundo no conjunto de dados DDR nas etapas de validação e teste da abordagem é apresentado na Tabela 32. Os melhores resultados foram obtidos com a utilização de *Tilling* e otimizador SGD, alcançando nas etapas de validação e teste tempo médio de inferência de 170ms e 190ms, respectivamente, conforme indicado em negrito.

Tabela 32 – Tempo médio de inferência para detectar as lesões de fundo no conjunto de dados DDR nas etapas de validação e teste com a abordagem proposta para segmentação de instância.

Modelos	Tempo de Inferência (ms)	
	Validação	**Teste**
Abordagem para Segmentação de Instância+SGD sem *Tilling*	510	700
Abordagem para Segmentação de Instância+Adam sem *Tilling*	840	1180
Abordagem para Segmentação de Instância+SGD com *Tilling*	**170**	**190**
Abordagem para Segmentação de Instância+Adam com *Tilling*	200	240

Ainda, comparou-se a abordagem para segmentação de instância com trabalhos relacionados encontrados na literatura, como apresentado na Tabela 33. Os resultados alcançados no conjunto de validação do *dataset* DDR são apresentados por meio das métricas AP e mAP com limite de IoU de 0,5. O melhor resultado de mAP obtido pela abordagem na tarefa de detecção (BBox) das lesões foi de 0,2903 (indicado em negrito), inferior ao resultado apresentado pelo trabalho de Shenavarmasouleh et al. (2021), que apresentou mAP de 0,4780 (indicado em negrito). Entretanto, convém destacar que Shenavarmasouleh et al. (2021) limitou-se a detectar somente duas classes de lesões: exsudatos e microaneurismas.

Tabela 33 – Resultados obtidos pela abordagem na detecção e segmentação das lesões de fundo em comparação a trabalhos relacionados encontrados na literatura utilizando a métrica AP e o limite de IoU de 0,5 no conjunto de dados DDR.

Modelos	**Tarefa**	*AP*				*mAP*
		EX	**SE**	**MA**	**HE**	
DeepLab-v3+ (LI et al., 2019)	BBox	-	-	-	-	-
	Mask	0,4375	0,4451	0,0390	0,4053	**0,3317**
HED (LI et al., 2019)	BBox	-	-	-	-	-
	Mask	0,2709	0,1612	0,0600	0,1885	0,1702
Mask R-CNN (SHENAVARMASOULEH et al., 2021)	BBox	-	-	-	-	**0,4780**
	Mask	-	-	-	-	-
Abordagem para Segmentação de Instância sem Tilling	BBox	0,2598	0,3147	0,1087	0,2585	0,2354
	Mask	0,2467	0,3400	0,1190	0,2342	0,2350
Abordagem para Segmentação de Instância com Tilling	BBox	0,2941	0,3508	0,1644	0,3520	**0,2903**
	Mask	0,3012	0,3505	0,2036	0,3211	**0,2941**

Isto possivelmente ocasionou um viés de classificação na classe majoritária, devido ao desbalanceamento do número de exemplos destas duas classes de lesões, motivo pelo qual o mAP apresentado possui valor elevado se comparado a trabalhos similares encontrados na literatura. Os autores também não apresentam os resultados

de AP por lesão a fim de se verificar a precisão obtida pelo modelo na detecção das microlesões. No trabalho também não constam os resultados obtidos na segmentação das lesões investigadas pelos autores.

Na tarefa de segmentação o trabalho proposto por Li et al. (2019) apresentou mAP de 0,3317 (indicado em negrito), enquanto que o melhor resultado obtido pela abordagem foi mAP igual a 0,2941 (indicado em negrito). Embora o resultado obtido para mAP seja menor, é fundamental destacar que a abordagem proposta apresenta uma melhoria na detecção de Microaneurismas, chegando a um AP aproximadamente 16% superior, significativamente melhor quando comparado ao de Li et al. (2019), para o qual o melhor valor de AP declarado foi de 0,0600. Além disso, a abordagem realizou a segmentação de instância, em que as lesões são detectadas e segmentadas, diferentemente do trabalho proposto por Li et al. (2019) no qual as lesões foram somente segmentadas.

Para entender melhor os resultados obtidos, na Figura 57 é apresentado um exemplo de segmentação de instância de lesões realizada pela abordagem em uma imagem de fundo do conjunto de dados DDR. Foi utilizado um valor de `NMS Testing Threshold` de 0,5. É possível verificar que foram obtidas as localizações de *pixel* a *pixel* das lesões detectadas, bem como as caixas delimitadoras de cada lesão. Esse tipo de abordagem pode auxiliar de forma mais efetiva no diagnóstico médico, pois é possível visualizar com maior clareza a extensão e o tamanho da lesão, ao invés de apenas detectar e traçar uma caixa delimitadora ao redor da lesão na imagem.

Percebe-se que a abordagem para segmentação de instância obteve resultados promissores na segmentação de instância das lesões de fundo investigadas, mesmo com a presença de múltiplas lesões com tamanhos e formas variáveis. Pelo fato de o modelo Mask R-CNN possuir um módulo RPN é possível realizar a extração de ROIs da imagem que tenham maior probabilidade de conter lesões de fundo. Os resultados experimentais demonstraram que a abordagem apresenta, por exemplo, maior precisão na detecção de Exsudatos Algodonosos e menor precisão na detecção de Microaneurismas.

A dificuldade na detecção dos microaneurismas ocorre em virtude do tamanho reduzido deste tipo de lesão. Durante o treinamento da rede neural profunda ocorre a dissipação de gradiente mais acentuada e, portanto, a perda de precisão associada às lesões que têm área de cobertura menor que 3 *pixels*, como no caso dos microaneurismas. Devido ao tamanho diminuto destas lesões, a extração de características destes artefatos fica comprometida, ocasionando baixa precisão em função do detector confundir estas microlesões com o fundo da imagem, resultando inclusive na geração de altas taxas de falsos negativos.

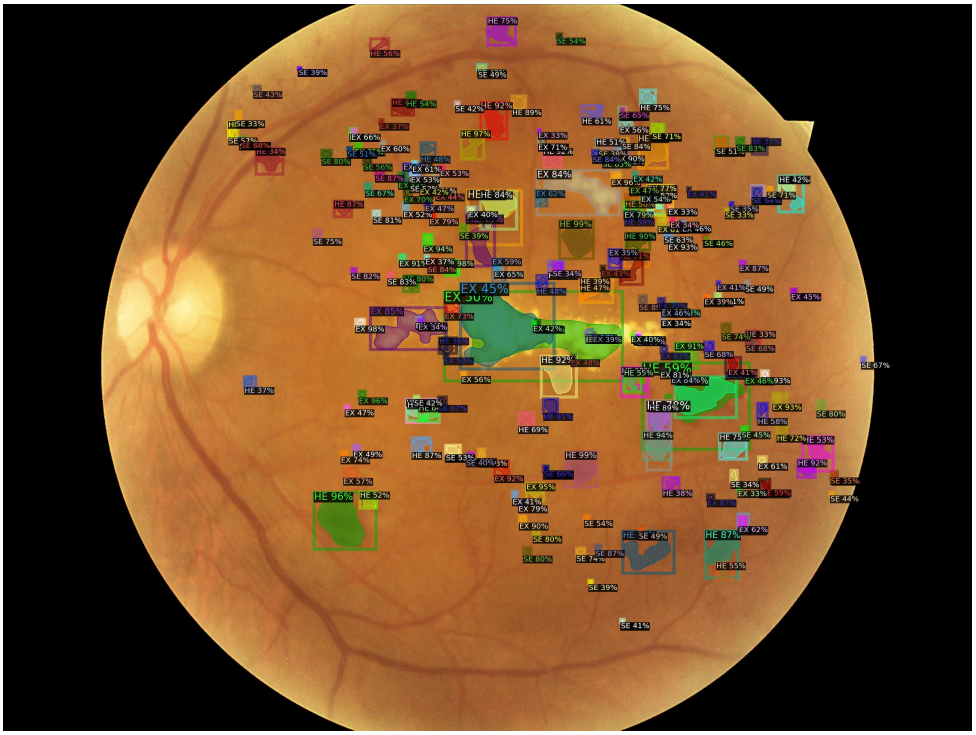


Figura 57 – Segmentação de instância de lesões de fundo realizada pela abordagem na imagem de fundo “007-3892-200.jpg” do conjunto de dados DDR. A classificação das lesões identificadas na imagem foi realizada em termos de *pixel*, sendo atribuído o rótulo da lesão e o percentual de confiança associado ao objeto detectado. Cada segmentação de instância tem uma cor diferente, independentemente da classe da lesão.

7.4 Comparação entre as Abordagens Propostas

A Tabela 34 apresenta os resultados de AP por lesão e mAP obtidos pelas abordagens propostas para detecção e segmentação de instância das lesões de fundo com otimizador Adam e *Tilling* durante a etapa de validação no conjunto de dados DDR. A Figura 58 apresenta o gráfico comparativo de AP por lesão obtido pelas abordagens propostas. Ambas abordagens alcançaram menor precisão na detecção dos Microaneurismas e maior precisão na detecção de Exsudatos Algodonosos e Hemorragias.

Tabela 34 – Resultados obtidos pelas abordagens propostas com *Tilling* e otimizador Adam com as métricas AP e mAP para o limite de IoU de 0,5 no conjunto de validação do conjunto de dados DDR.

Modelos	AP				mAP
	EX	**SE**	**MA**	**HE**	
Abordagem para Detecção	0,2240	**0,3650**	0,1110	**0,3520**	0,2630
Abordagem para Segmentação de Instância	**0,2941**	0,3508	**0,1644**	**0,3520**	**0,2903**

Os resultados obtidos decorrem do tamanho e das características morfológicas destas lesões. Os Microaneurismas, por exemplo, são microlesões com pouca cobertura de *pixels* que acaba ocasionando uma perda na precisão durante a detecção uma vez que estas microlesões são frequentemente confundidas com o fundo da imagem.

Além disso, é possível verificar que a abordagem para segmentação de instância foi 31,29% mais precisa na detecção dos Exsudatos duros, e 48,11% mais precisa na detecção de Microaneurismas em comparação à abordagem proposta para detecção. Entretanto, a abordagem para detecção foi 4,05% mais precisa na detecção de Exsudatos Algodonosos em relação à abordagem para segmentação de instância. No geral, a abordagem proposta para segmentação de instância foi 10,38% mais precisa, com mAP de 0,2903, em comparação à abordagem proposta para detecção, que atingiu mAP de 0,2630.

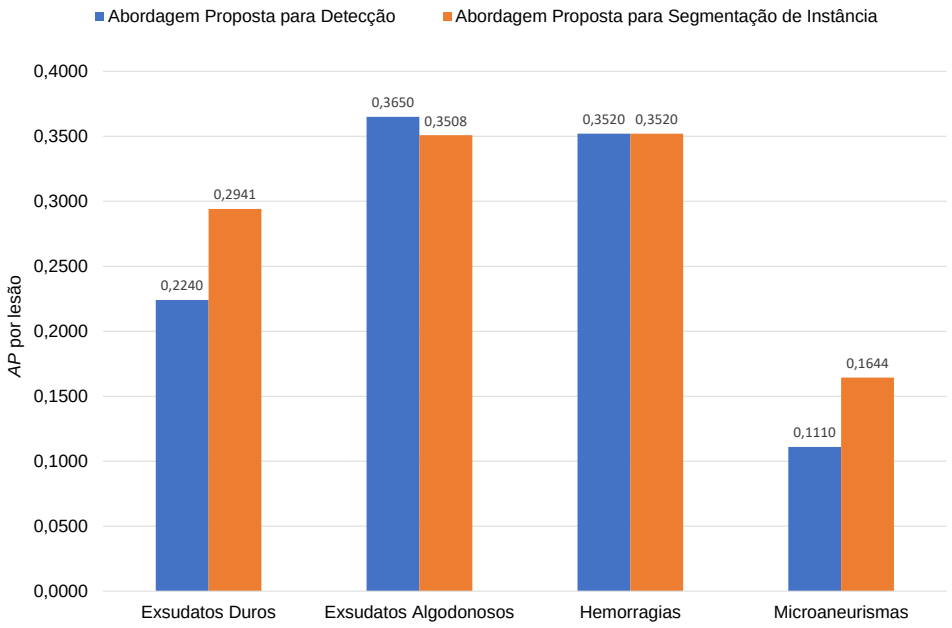


Figura 58 – Gráfico de AP por lesão das abordagens propostas para detecção e segmentação de instância das lesões de fundo associadas à RD com otimizador Adam e *Tilling* durante a etapa de validação no conjunto de dados DDR.

A Tabela 35 apresenta os tempos médios de inferência para detectar as lesões de fundo no conjunto de dados DDR durante as etapas de validação e teste obtidos pelas abordagens propostas com otimizador Adam e *Tilling*. Além disso, a Figura 59 apresenta o gráfico com os tempos médios de inferência em (ms) obtidos por ambas abordagens.

É possível observar que a abordagem para detecção realiza inferências mais rapidamente em comparação à abordagem para segmentação de instância, tendo alcançado tempo médio de inferência de 5,5 ms na etapa de validação e 7,5 ms na etapa de teste, ambos no conjunto de dados DDR e utilizando otimizador Adam. Em expe-

Tabela 35 – Tempo médio de inferência para detectar as lesões de fundo no conjunto de dados DDR nas etapas de validação e teste das abordagens propostas com otimizador Adam e *Tilling*.

Modelos	Tempo de Inferência (ms)	
	Validação	**Teste**
Abordagem para Detecção	**5,5**	**7,5**
Abordagem para Segmentação de Instância	200	240

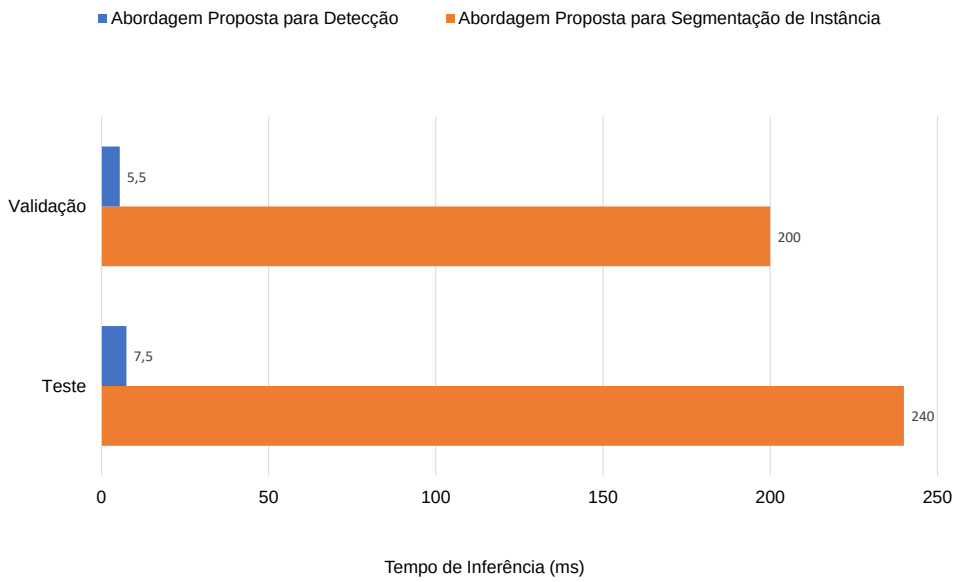


Figura 59 – Gráfico de tempo de inferência das abordagens propostas para detecção e segmentação de instância das lesões de fundo associadas à RD com otimizador Adam e *Tilling* durante as etapas de validação e teste no conjunto de dados DDR.

rimento análogo, a abordagem proposta para segmentação de instância obteve nas etapas de validação e teste tempos médios de inferência iguais a 200 ms e 240 ms, respectivamente. Embora a abordagem para segmentação de instância tenha sido mais precisa, a abordagem para detecção foi mais rápida na realização de predições, sendo aproximadamente 36 vezes mais rápida na etapa de validação e 32 vezes mais rápida na etapa de teste.

A abordagem proposta para detecção consegue manter um bom compromisso entre precisão e velocidade, ao passo que a abordagem proposta para segmentação de instância, embora seja relativamente mais precisa, tem um tempo de inferência elevado se comparado à abordagem para detecção.

Contudo, a segmentação de instância apresenta a vantagem de realizar a detecção e segmentação das lesões que pode ser mais útil para o diagnóstico médico das lesões retinianas. Em síntese, para a escolha de qual das abordagens propostas é mais adequada para o auxílio ao diagnóstico médico é necessário que seja levado

em conta este *trade-off* entre precisão e velocidade na escolha da abordagem mais adequada para compor uma aplicação destinada a realizar a detecção de lesões de fundo e auxiliar no diagnóstico médico.

7.5 Considerações sobre o Capítulo

Este Capítulo apresentou a abordagem para a segmentação de instância das lesões de fundo do olho. Foram mostradas as etapas de pré-processamento e aumento de dados, assim como a constituição da arquitetura de rede neural profunda utilizada.

Os conjuntos de dados públicos de Retinopatia Diabética utilizados nos experimentos foram particionados em uma proporção de 50:20:30 para aplicar o processo de treinamento, validação e teste, respectivamente. Usou-se uma arquitetura de rede neural profunda Mask R-CNN implementada por meio da biblioteca Detectron2.

Também foram apresentados os cenários experimentais, testes e avaliação dos resultados obtidos pela abordagem. Na sequência, foi realizada uma discussão sobre as abordagens propostas para detecção e segmentação de instância das lesões de fundo associadas à retinopatia diabética com o objetivo de comparar os principais resultados obtidos por ambas as abordagens.

O próximo Capítulo contempla as considerações finais deste trabalho, sendo discutidas as principais conclusões atingidas. Também são relacionadas as publicações realizadas durante o doutorado, bem como são destacados alguns potenciais trabalhos futuros relacionados à Tese.

8 CONSIDERAÇÕES FINAIS

Esta Tese propôs duas novas abordagens usando diferentes técnicas de processamento de imagens e arquiteturas de redes neurais profundas para aumentar a precisão da detecção de lesões de fundo associadas à Retinopatia Diabética. Esta Tese foi motivada pelo crescimento do Diabetes e Retinopatia Diabética na população mundial e, também, em função dos custos envolvidos com profissionais da área da saúde e demais recursos necessários para a triagem, acompanhamento e tratamento destas enfermidades, e a necessidade de criação de soluções que possam dar um suporte mais rápido e preciso no diagnóstico médico, a fim de proporcionar uma avaliação e tratamento mais adequados para um número maior de pessoas.

Como primeira contribuição desta Tese, apresenta-se uma nova abordagem para a detecção de lesões de fundo explicada no Capítulo 6. Os resultados obtidos pela abordagem proposta para detecção apresentam uma precisão superior a trabalhos de mesmo propósito apresentados na literatura. Além disso, a análise dos resultados permitiu identificar as lesões que apresentaram menores taxas de detecção, o que foi determinante para o desenvolvimento das soluções utilizadas para aumentar a precisão na detecção destas lesões.

Em relação à abordagem para a detecção das lesões de fundo foi implementado um bloco de pré-processamento para reduzir *outliers* e melhorar o realce, a fim de prover uma extração de características mais eficiente das lesões retinianas. Este bloco conta com um método para *Cropping* parcial do fundo preto das imagens de fundo, para minimizar a geração de falsos positivos, e também um método de *Tilling*, para aumentar o campo receptivo em torno das lesões e minimizar a perda de informação causada pela redução de resolução das imagens originais na camada de entrada do *Backbone* da arquitetura. Além disso, foi apresentada uma estrutura de rede neural convolucional baseada em uma arquitetura YOLO de última geração para melhorar a detecção das lesões de fundo e realizar inferências em tempo real sobre GPUs de baixo custo.

A solução proposta para a detecção das lesões de fundo foi treinada e avaliada utilizando dois conjuntos de dados públicos de Retinopatia Diabética: DDR e IDRiD. Particionou-se estes conjuntos de dados em conjunto de treinamento, validação e teste em uma proporção de 50:20:30, respectivamente. Os melhores resultados alcançados por esta abordagem no conjunto de dados DDR foram obtidos utilizando o otimizador Adam e o método de *Tilling*, alcançando para o limite de IoU de 0,5 na etapa de validação mAP de 0,2630, e na etapa de teste mAP de 0,1540.

Como segunda contribuição desta Tese, apresentou-se uma nova abordagem para segmentação de instância de lesões de fundo apresentada no Capítulo 7. Os resultados alcançados pela abordagem para realizar a segmentação de instância apresenta uma precisão superior à abordagem proposta para detecção. Também apresenta resultados superiores na identificação de microaneurismas em relação a trabalhos de propósito similar relatados na literatura. Ainda, em relação à abordagem para a segmentação de instância das lesões de fundo, apresenta-se uma estrutura de rede neural convolucional implementada com base em uma arquitetura Mask R-CNN para melhorar a precisão na detecção das lesões de fundo, principalmente na identificação das microlesões. A abordagem para segmentação de instância também contém um bloco de pré-processamento composto pelos métodos de *Cropping* parcial do fundo preto das imagens de fundo e *Tilling* para a criação de sub-imagens a partir das imagens originais dos conjuntos de dados públicos de RD para treinamento da rede neural.

A abordagem para segmentação de instância também foi treinada e avaliada utilizando os conjuntos de dados DDR e IDRiD, e particionados em conjunto de treinamento, validação e teste na proporção de 50:20:30, respectivamente. Os melhores resultados alcançados por esta abordagem no conjunto de dados DDR foram obtidos utilizando o otimizador Adam e o método de *Tilling*, atingindo para o limite de IoU de 0,5 na etapa de validação mAP de 0,2903, e na etapa de teste mAP de 0,1670.

As soluções propostas nesta Tese apresentaram resultados competitivos em relação aos trabalhos relacionados. Os resultados obtidos nos experimentos demonstram que a abordagem para detecção de lesões apresentou resultados superiores a trabalhos de mesmo propósito encontrados na literatura. No geral, a abordagem para segmentação de instância de lesões foi 10,38% mais precisa na detecção das lesões em relação à abordagem para detecção, sendo 31,29% mais precisa na detecção dos Exsudatos Duros e 48,11% na detecção dos Microaneurismas.

Embora a abordagem para segmentação de instância seja mais precisa, a abordagem proposta para detecção é mais rápida na realização das predições. A abordagem para detecção consegue manter um bom compromisso entre precisão e velocidade, ao passo que a abordagem proposta para segmentação de instância, embora seja relativamente mais precisa, tem um tempo de inferência elevado se comparado à abordagem para detecção. Contudo, a segmentação de instância apresenta a van-

tagem de realizar a detecção e segmentação das lesões que pode ser mais útil para o diagnóstico médico das lesões retinianas.

Resumindo, nesta Tese foram desenvolvidas duas novas abordagens para detecção e segmentação de instância de lesões de fundo usando técnicas de processamento de imagem e redes neurais profundas implementadas com base em detectores de objetos de última geração. As avaliações demonstraram que as abordagens propostas apresentaram resultados promissores na detecção das lesões retinianas associadas à Retinopatia Diabética investigadas. Entretanto, há um vasto espaço que ainda pode ser explorado para melhorar a precisão na identificação destas lesões em imagens de fundo.

Como trabalho futuro, pretende-se desenvolver soluções que combinem diferentes contextos, como no aprendizado multimodal a fim de obter resultados mais precisos, e proporcionar ao médico especialista uma quantidade maior de informações para auxiliá-lo na realização do diagnóstico. Além disso, embora a solução proposta neste trabalho forneça resultados significativos na detecção das lesões de fundo, há muitos pontos que podem ser considerados para pesquisas futuras. Por exemplo, a utilização de modelos de difusão para sintetizar novas imagens juntamente com as anotações das lesões, o desenvolvimento de modelos que contenham mecanismos de atenção para aumentar a precisão da detecção das microlesões, ou até mesmo a utilização de arquiteturas de redes neurais profundas com foco na detecção de objetos sem a utilização de âncoras. Portanto, há um vasto espaço para explorar soluções a fim de melhorar a detecção das lesões retinianas e auxiliar no diagnóstico precoce da RD.

8.1 Contribuições Científicas

A pesquisa realizada durante esta Tese de Doutorado resultou na publicação de nove artigos científicos completos, sendo três artigos com Qualis A1, dois artigos com Qualis A2, um artigo com Qualis A3, e dois artigos com Qualis A4. A seguir são apresentadas as informações destes artigos publicados como primeiro autor e coautor:

Pôster

Título:	Detection of Fundus of the Eye Injuries Through Convolutional Neural Networks
Autores:	Carlos Santos, Marilton Aguiar, Daniel Welfer
Evento:	4th IEEE Seasonal School on Digital Processing of Visual Signals and Applications (IEEE DPVSA)
Ano:	2020
Link:	`https://wp.ufpel.edu.br/dpvsa`

Artigos completos como primeiro autor

Título: Deep Neural Network Model based on One-Stage Detector for Identifying Fundus Lesions
Autores: Carlos Santos, Marilton Aguiar, Daniel Welfer, Bruno Belloni
Conferência: International Joint Conference on Neural Networks (IJCNN)
Ano: 2021
DOI: https://doi.org/10.1109/IJCNN52387.2021.9534354
Qualis: A1

Título: Detection of Fundus Lesions through a Convolutional Neural Network in Patients with Diabetic Retinopathy
Autores: Carlos Santos, Marilton Aguiar, Daniel Welfer, Bruno Belloni
Conferência: 43rd Annual International Conference of the IEEE Engineering in Medicine and Biology Society (IEEE EMBC)
Ano: 2021
DOI: https://doi.org/10.1109/EMBC46164.2021.9630075
Qualis: A1

Título: A New Method Based on Deep Learning to Detect Lesions in Retinal Images using YOLOv5
Autores: Carlos Santos, Marilton Aguiar, Daniel Welfer, Bruno Belloni
Conferência: IEEE International Conference on Bioinformatics and Biomedicine (IEEE BIBM)
Ano: 2021
DOI: https://doi.org/10.1109/BIBM52615.2021.9669581
Qualis: A2

Título: A New Approach for Detecting Fundus Lesion using Image Processing and Deep Neural Network Architecture based on YOLO model
Autores: Carlos Santos, Marilton Aguiar, Daniel Welfer, Bruno Belloni
Periódico: Sensors
Ano: 2022
DOI: https://doi.org/10.3390/s22176441
Qualis: A2

Título: A Method based on Deep Neural Network for Instance Segmentation of Retinal Lesions caused by Diabetic Retinopathy
Autores: Carlos Santos, Marilton Aguiar, Daniel Welfer, Marcelo Silva, Alejandro Pereira, Marcelo Ribeiro, Bruno Belloni
Conferência: 9th Annual Conference on Computational Science and Computational Intelligence (CSCI)
Ano: 2022
DOI: https://doi.org/10.1109/CSCI58124.2022.00033
Qualis: A4

Título: A New Approach for Fundus Lesions Instance Segmentation Based on Mask R-CNN X101-FPN Pre-Trained Architecture
Autores: Carlos Santos, Marilton Aguiar, Daniel Welfer, Marcelo Dias, Alejandro Pereira, Marcelo Ribeiro, Bruno Belloni
Periódico: IEEE Access
Ano: 2023
DOI: https://doi.org/10.1109/ACCESS.2023.3271895
Qualis: A3

Artigos completos como coautor

Título: Detection of retinal microlesions through YOLOR-CSP architecture and image slicing with the SAHI algorithm
Autores: Alejandro Pereira, Carlos Santos, Marilton Aguiar, Daniel Welfer, Marcelo Dias, Marcelo Ribeiro, Reza Ahmadi
Conferência: International Joint Conference on Neural Networks (IJCNN)
Ano: 2023
DOI: https://doi.org/10.1109/IJCNN54540.2023.10191623
Qualis: A1

Título: Improved Detection of Fundus Lesions Using YOLOR-CSP Architecture and Slicing Aided Hyper Inference
Autores: Alejandro Pereira, Carlos Santos, Marilton Aguiar, Daniel Welfer, Marcelo Dias, Marcelo Ribeiro
Periódico: IEEE Latin America Transactions
Ano: 2023
DOI: http://dx.doi.org/10.1109/tla.2023.10244179
Qualis: B2

Título: Um Novo Método baseado em Detector de Dois Estágios para Segmentação de Instância de Lesões Retinianas usando o Modelo Mask R-CNN e a Biblioteca Detectron2
Autores: Marcelo Dias, Carlos Santos, Marilton Aguiar, Daniel Welfer, Alejandro Pereira, Marcelo Ribeiro
Conferência: 50º Seminário Integrado de Software e Hardware (SEMISH)
Ano: 2023
DOI: https://doi.org/10.5753/semish.2023
Qualis: A4

REFERÊNCIAS

AACH, T.; KAUP, A.; MESTER, R. On texture analysis: Local energy transforms versus quadrature filters. **Signal Processing**, Amesterdã, v.45, n.2, p.173–181, 1995.

AGARAP, A. F. **Deep Learning using Rectified Linear Units (ReLU)**.

AGARWAL, N.; ANIL, R.; HAZAN, E.; KOREN, T.; ZHANG, C. **Revisiting the Generalization of Adaptive Gradient Methods**. Disponível em: <https://openreview.net/forum?id=BJl6t64tvr>.

AHMAD FADZIL, M. H.; IZHAR, L. I.; NUGROHO, H.; NUGROHO, H. A. Analysis of retinal fundus images for grading of diabetic retinopathy severity. **Medical and Biological Engineering and Computing**, Berlin, v.49, n.6, p.693–700, 2011.

AHONEN, T.; HADID, A.; PIETIKAINEN, M. Face Description with Local Binary Patterns: Application to Face Recognition. **IEEE Transactions on Pattern Analysis and Machine Intelligence**, Piscataway, New Jersey, v.28, n.12, p.2037–2041, 2006.

AIDOUNI, M. E. **Evaluating Object Detection Models**: Guide to Performance Metrics. `https://manalelaidouni.github.io/Evaluating-Object-Detection-Models-Guide-to-Performance-Metrics.html`.

AKYOL, G.; KANTARCı, A.; ÇELIK, A. E.; CIHAN AK, A. Deep Learning Based, Real-Time Object Detection for Autonomous Driving. In: SIGNAL PROCESSING AND COMMUNICATIONS APPLICATIONS CONFERENCE (SIU), 28., 2020, Gaziantep, Turkey. **Proceedings...** IEEE, 2020. p.1–4.

AKYOL, K.; SEN, B.; BAYIR, S. Automatic Detection of Optic Disc in Retinal Image by Using Keypoint Detection, Texture Analysis, and Visual Dictionary Techniques. **Computational and Mathematical Methods in Medicine**, London, UK, v.2016, 2016.

AL-BANDER, B.; AL-NUAIMY, W.; WILLIAMS, B. M.; ZHENG, Y. Multiscale sequential convolutional neural networks for simultaneous detection of fovea and optic disc. **Biomedical Signal Processing and Control**, Amesterdã, v.40, p.91–101, 2018.

AL-KADI, O. S. **A Gabor Filter Texture Analysis Approach for Histopathological Brain Tumor Subtype Discrimination**.

AL-MASNI, M. A.; AL-ANTARI, M. A.; PARK, J. M.; GI, G.; KIM, T. Y.; RIVERA, P.; VALAREZO, E.; CHOI, M. T.; HAN, S. M.; KIM, T. S. Simultaneous detection and classification of breast masses in digital mammograms via a deep learning YOLO-based CAD system. **Computer Methods and Programs in Biomedicine**, Amesterdã, v.157, p.85–94, 2018.

ALGHADYAN, A. A. Diabetic retinopathy - An update. **Saudi Journal of Ophthalmology**, Amesterdã, v.25, n.2, p.99–111, 2011.

ALOYSIUS, N.; GEETHA, M. A review on deep convolutional neural networks. In: INTERNATIONAL CONFERENCE ON COMMUNICATION AND SIGNAL PROCESSING (ICCSP), 2017, Chennai, India. **Proceedings...** IEEE, 2017. p.0588–0592.

ALYOUBI, W. L.; ABULKHAIR, M. F.; SHALASH, W. M. Diabetic Retinopathy Fundus Image Classification and Lesions Localization System Using Deep Learning. **Sensors**, Basel, Switzerland, v.21, n.11, 2021.

AMERIKANOS, P.; MAGLOGIANNIS, I. Image Analysis in Digital Pathology Utilizing Machine Learning and Deep Neural Networks. **Journal of Personalized Medicine**, Basel, Switzerland, v.12, n.9, 2022.

AMORIM, J. G. **Como calcular Average Precision - AP e Intersection over Union - IoU**. `https://gist.github.com/johnnv1/fdbd901428daebe2be0402f7ab75fb17`.

ANDREWS, B. W.; POLLEN, D. A. Relationship between spatial frequency selectivity and receptive field profile of simple cells. **The Journal of Physiology**, Hoboken, New Jersey, v.287, n.1, p.163–176, 1979.

ARNAB, A.; TORR, P. H. S. **Pixelwise Instance Segmentation with a Dynamically Instantiated Network**.

BENZAMIN, A.; CHAKRABORTY, C. Detection of Hard Exudates in Retinal Fundus Images Using Deep Learning. In: IEEE INTERNATIONAL CONFERENCE ON SYSTEM, COMPUTATION, AUTOMATION AND NETWORKING, ICSCA 2018, 2018, Pondicherry, India. **Proceedings...** IEEE, 2018.

BERTELS, J.; EELBODE, T.; BERMAN, M.; VANDERMEULEN, D.; MAES, F.; BISSCHOPS, R.; BLASCHKO, M. B. Optimizing the Dice Score and Jaccard Index for Medical Image Segmentation: Theory and Practice. **Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)**, Berlin, v.11765 LNCS, p.92–100, 2019.

BESENCZI, R.; TÓTH, J.; HAJDU, A. A review on automatic analysis techniques for color fundus photographs. **Computational and Structural Biotechnology Journal**, Amesterdã, v.14, p.371–384, 2016.

BISHOP, B. Y. P. O.; COOMBS, J. S.; HENRY, G. H. Receptive fields of simple cells in the cat striate cortex. **J Physiol.**, Bethesda, Maryland, p.31–60, 1973.

BLITZER, J.; DREDZE, M.; PEREIRA, F. Biographies, Bollywood, Boom-boxes and Blenders: Domain Adaptation for Sentiment Classification. In: ANNUAL MEETING OF THE ASSOCIATION OF COMPUTATIONAL LINGUISTICS, 45., 2007, Prague, Czech Republic. **Proceedings...** Association for Computational Linguistics, 2007. p.440–447.

BOCHKOVSKIY, A.; WANG, C.-Y.; LIAO, H.-Y. M. YOLOv4: Optimal Speed and Accuracy of Object Detection. **ArXiv e-prints**, New York, NY, 2020. arXiv:2004.10934.

BODLA, N.; SINGH, B.; CHELLAPPA, R.; DAVIS, L. S. Soft-NMS - Improving Object Detection with One Line of Code. In: IEEE INTERNATIONAL CONFERENCE ON COMPUTER VISION, 2017, Venice, Italy. **Proceedings...** IEEE, 2017. v.2017-October, p.5562–5570.

BUDA, M.; MAKI, A.; MAZUROWSKI, M. A. A systematic study of the class imbalance problem in convolutional neural networks. **Neural Networks**, Amesterdã, v.106, p.249–259, Oct 2018.

BUSLAEV, A.; IGLOVIKOV, V. I.; KHVEDCHENYA, E.; PARINOV, A.; DRUZHININ, M.; KALININ, A. A. Albumentations: Fast and flexible image augmentations. **Information (Switzerland)**, Basel, Switzerland, v.11, n.2, 2020.

CARDOSO, C. d. F. d. S. **Segmentação automática do disco óptico e de vasos sanguíneos em imagens de fundo de olho**. 2019. 149p. Tese de Doutorado (Curso de Doutorado em Ciências) — Universidade Federal de Uberlândia, Uberlândia.

CARRATINO, L.; CISSé, M.; JENATTON, R.; VERT, J.-P. **On Mixup Regularization**.

CASTRO, C. L. de; BRAGA, A. P. Aprendizado supervisionado com conjuntos de dados desbalanceados. **Sba: Controle & Automação Sociedade Brasileira de Automática**, São Paulo, SP, v.22, n.5, p.441–466, 2011.

CASTRO, D. J. L. **Garra servo-controlada com integração de informação táctil e de proximidade**. 1996. 133p. Dissertação de Mestrado (Curso de Mestrado em Engenharia Eletrotécnica) — Universidade de Coimbra, Coimbra.

CHAKRABARTI, R.; HARPER, C. A.; KEEFFE, J. E. Diabetic retinopathy management guidelines. **Expert Review of Ophthalmology**, London, UK, v.7, n.5, p.417–439, 2012.

CHANDRASEKAR, L.; DURGA, G. Implementation of Hough Transform for image processing applications. In: INTERNATIONAL CONFERENCE ON COMMUNICATION AND SIGNAL PROCESSING, 2014, Melmaruvathur, India. **Proceedings...** IEEE, 2014. p.843–847.

CHEN, J.; WOLFE, C.; LI, Z.; KYRILLIDIS, A. **Demon**: Improved Neural Network Training with Momentum Decay.

CHEN, K. et al. MMDetection: Open MMLab Detection Toolbox and Benchmark. **arXiv preprint arXiv:1906.07155**, New York, NY, 2019.

CHEN, L. C.; ZHU, Y.; PAPANDREOU, G.; SCHROFF, F.; ADAM, H. Encoder-decoder with atrous separable convolution for semantic image segmentation. **Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)**, Berlin, v.11211 LNCS, p.833–851, 2018.

CHOI, D.; SHALLUE, C. J.; NADO, Z.; LEE, J.; MADDISON, C. J.; DAHL, G. E. **On Empirical Comparisons of Optimizers for Deep Learning**.

CHORIANOPOULOS, A. M.; DARAMOUSKAS, I.; PERIKOS, I.; GRIVOKOSTOPOULOU, F.; HATZILYGEROUDIS, I. Deep Learning Methods in Medical Imaging for the Recognition of Breast Cancer. In: INTERNATIONAL CONFERENCE ON INFORMATION, INTELLIGENCE, SYSTEMS AND APPLICATIONS (IISA), 11., 2020, Piraeus, Greece. **Proceedings...** IEEE, 2020. p.1–8.

CHUDZIK, P.; AL-DIRI, B.; CALIVA, F.; OMETTO, G.; HUNTER, A. Exudates Segmentation using Fully Convolutional Neural Network and Auxiliary Codebook. In: ANNUAL INTERNATIONAL CONFERENCE OF THE IEEE ENGINEERING IN MEDICINE AND BIOLOGY SOCIETY, EMBS, 2018, Honolulu, USA. **Proceedings...** IEEE, 2018. v.2018-July, p.770–773.

CLARO, M.; VOGADO, L.; SANTOS, J.; VERAS, R. **Utilização de Técnicas de Data Augmentation em Imagens: Teoria e Prática**. Online; accessed 01-Nov-2021, `https://sol.sbc.org.br/livros/index.php/sbc/catalog/view/48/224/445-1`.

COCO. **Detection evaluation metrics used by COCO**. `https://cocodataset.org/#detection-eval`.

COUTURIER, R.; NOURA, H. N.; SALMAN, O.; SIDER, A. **A Deep Learning Object Detection Method for an Efficient Clusters Initialization**.

D. D. Silva, A.; B. P. Carneiro, M.; F. S. Cardoso, C. Realce De Microaneurimas Em Imagens De Fundo De Olho Utilizando Clahe. In: V CONGRESSO BRASILEIRO DE ELETROMIOGRAFIA E CINESIOLOGIA E X SIMPóSIO DE ENGENHARIA BIOMéDICA, 2018, Uberlândia. **Anais...** Even3, 2018. p.772–775.

DAI, F.; FAN, B.; PENG, Y. An image haze removal algorithm based on blockwise processing using LAB color space and bilateral filtering. In: CHINESE CONTROL AND DECISION CONFERENCE (CCDC), 2018. **Proceedings...** IEEE, 2018. p.5945–5948.

DAI, J.; HE, K.; SUN, J. **Instance-aware Semantic Segmentation via Multi-task Network Cascades**.

DAI, J.; LI, Y.; HE, K.; SUN, J. **R-FCN**: Object Detection via Region-based Fully Convolutional Networks.

DAI, J.; QI, H.; XIONG, Y.; LI, Y.; ZHANG, G.; HU, H.; WEI, Y. Deformable Convolutional Networks. In: IEEE INTERNATIONAL CONFERENCE ON COMPUTER VISION, 2017, Venice, Italy. **Proceedings...** IEEE, 2017. v.2017-Octob, p.764–773.

DAI, L.; WU, L.; LI, H.; CAI, C.; WU, Q.; KONG, H.; LIU, R.; WANG, X.; HOU, X.; LIU, Y.; LONG, X.; WEN, Y.; LU, L.; SHEN, Y.; CHEN, Y.; SHEN, D.; YANG, X.; ZOU, H.; SHENG, B.; JIA, W. A deep learning system for detecting diabetic retinopathy across the disease spectrum. **Nature Communications**, Berlin, v.12, n.1, 2021.

DALAL, N.; TRIGGS, B. Histograms of oriented gradients for human detection. In: IEEE COMPUTER SOCIETY CONFERENCE ON COMPUTER VISION AND PATTERN RECOGNITION (CVPR'05), 2005, San Diego, California. **Proceedings...** IEEE, 2005. v.1, p.886–893 vol. 1.

DAVIS, J.; GOADRICH, M. The relationship between Precision-Recall and ROC curves. In: INTERNATIONAL CONFERENCE ON MACHINE LEARNING, 23., 2006, Pittsburgh, Pennsylvania. **Proceedings...** ACM, 2006. v.2006, p.233–240.

DELGADO-BONAL, A.; MARTÍN-TORRES, J. Human vision is determined based on information theory. **Scientific Reports**, Berlin, v.6, n.October, p.1–5, 2016.

DEWI, C.; CHEN, R. C.; LIU, Y. T.; JIANG, X.; HARTOMO, K. D. Yolo V4 for Advanced Traffic Sign Recognition with Synthetic Training Data Generated by Various GAN. **IEEE Access**, Piscataway, New Jersey, v.9, p.97228–97242, 2021.

DORION, T. **Manual de Exame Do Fundo de Olho**. 1.ed. Barueri, SP: Manole, 2002.

DUCHI, J.; HAZAN, E.; SINGER, Y. Adaptive Subgradient Methods for Online Learning and Stochastic Optimization. **J. Mach. Learn. Res.**, New York, NY, v.12, n.null, p.2121–2159, July 2011.

DVORNIK, N.; MAIRAL, J.; SCHMID, C. **Modeling Visual Context is Key to Augmenting Object Detection Datasets**.

DWIBEDI, D.; MISRA, I.; HEBERT, M. **Cut, Paste and Learn**: Surprisingly Easy Synthesis for Instance Detection.

EL ABBADI, N.; HAMMOD, E. Automatic Early Diagnosis of Diabetic Retinopathy Using Retina Fundus Images Enas Hamood Al-Saadi-Automatic Early Diagnosis of Diabetic Retinopathy Using Retina Fundus Images. **EAR**, Beijing, v.2, 12 2014.

ELFWING, S.; UCHIBE, E.; DOYA, K. **Sigmoid-Weighted Linear Units for Neural Network Function Approximation in Reinforcement Learning**.

ELHEDDA, W.; MEHRI, M. A Comparative Study of Filtering Approaches Applied to Color Archival Document Images. **CoRR**, New York, NY, v.abs/1908.0, p.1–8, 2019.

ELSAYED, A. A.; YOUSEF, W. A. **Matlab vs. OpenCV**: A Comparative Study of Different Machine Learning Algorithms. New York, NY: Arxiv, 2019. Disponível em: <https://arxiv.org/abs/1905.01213>.

ERFURT, J.; HELMRICH, C. R.; BOSSE, S.; SCHWARZ, H.; MARPE, D.; WIEGAND, T. A Study of the Perceptually Weighted Peak Signal-To-Noise Ratio (WPSNR) for Image Compression. In: IEEE INTERNATIONAL CONFERENCE ON IMAGE PROCESSING (ICIP), 2019, Taipei, Taiwan. **Proceedings...** IEEE, 2019. p.2339–2343.

ETDRSR, E. T. D. R. S. R. G. Grading Diabetic Retinopathy from Stereoscopic Color Fundus Photographs—An Extension of the Modified Airlie House Classification. **Ophthalmology**, Amesterdã, v.98, n.5, p.786–806, May 1991.

ETDRSR, E. T. D. R. S. R. G. Classification of Diabetic Retinopathy from Fluorescein Angiograms. **Ophthalmology**, Amesterdã, v.98, n.5, p.807–822, May 1991.

FARDO, F. A.; CONFORTO, V. H.; OLIVEIRA, F. C. de; RODRIGUES, P. S. **A Formal Evaluation of PSNR as Quality Measurement Parameter for Image Segmentation Algorithms**.

FARIA, D. **Trabalhos Práticos Análise e Processamento de Imagem**. Disponível em: <https://bit.ly/3CV8Oi3>.

FAUST, O.; Acharya U., R.; NG, E. Y.; NG, K. H.; SURI, J. S. Algorithms for the automated detection of diabetic retinopathy using digital fundus images: A review. **Journal of Medical Systems**, Berlin, v.36, n.1, p.145–157, 2012.

FEI-FEI, L.; DENG, J.; LI, K. ImageNet: Constructing a large-scale image database. **Journal of Vision**, Washington, DC, v.9, n.8, p.1037–1037, 2010.

FERNÁNDEZ, A.; GARCÍA, S.; GALAR, M.; PRATI, R. C. **Learning from Imbalanced Data Sets**. Berlin, Germany: Springer, 2019. 1–377p.

FLACH, P. A.; KULL, M. Precision-Recall-Gain curves: PR analysis done right. **Advances in Neural Information Processing Systems**, Montreal, Quebec, v.2015-January, p.838–846, 2015.

FRANKE, M.; GOPINATH, V.; REDDY, C.; RISTIĆ-DURRANT, D.; MICHELS, K. Bounding Box Dataset Augmentation for Long-Range Object Distance Estimation. In: IEEE/CVF INTERNATIONAL CONFERENCE ON COMPUTER VISION (ICCV) WORKSHOPS, 2021, Virtual Conference. **Proceedings...** IEEE, 2021. p.1669–1677.

FREITAS, G. A. d. L. **Aprendizagem Profunda Aplicada ao Futebol de Robôs** : Uso de Redes Neurais Convolucionais para Detecção de Objetos Universidade Estadual de Londrina Centro de Tecnologia e Urbanismo Departamento de Engenharia Elétrica Aprendizagem Profunda Aplicada ao Fute. 2019. 56p. Trabalho de Conclusão (Curso de Engenharia Elétrica) — Universidade Estadual de Londrina, Londrina.

FUAD, M. T. H.; FIME, A. A.; SIKDER, D.; IFTEE, M. A. R.; RABBI, J.; AL-RAKHAMI, M. S.; GUMAEI, A.; SEN, O.; FUAD, M.; ISLAM, M. N. Recent Advances in Deep Learning Techniques for Face Recognition. **IEEE Access**, Piscataway, New Jersey, v.9, p.99112–99142, 2021.

GAGNON, L.; LALONDE, M.; BEAULIEU, M.; BOUCHER, M.-C. Procedure to detect anatomical structures in optical fundus images. **Medical Imaging 2001: Image Processing**, [S.l.], v.4322, p.1218–1225, 2001.

GANESH, P. **Object Detection : Simplified**. [Online; accessed 12-June-2021], `https://towardsdatascience.com/object-detection-simplified-e07aa3830954`.

GE, Z.; LIU, S.; WANG, F.; LI, Z.; SUN, J. **YOLOX**: Exceeding YOLO Series in 2021.

GHIASI, G.; CUI, Y.; SRINIVAS, A.; QIAN, R.; LIN, T.-Y.; CUBUK, E. D.; LE, Q. V.; ZOPH, B. **Simple Copy-Paste is a Strong Data Augmentation Method for Instance Segmentation**.

GIRSHICK, R. Fast R-CNN. In: IEEE INTERNATIONAL CONFERENCE ON COMPUTER VISION, 2015, Santiago, Chile. **Proceedings...** IEEE, 2015. v.2015 International Conference on Computer Vision, ICCV 2015, p.1440–1448.

GIRSHICK, R.; DONAHUE, J.; DARRELL, T.; MALIK, J. Rich feature hierarchies for accurate object detection and semantic segmentation. In: IEEE COMPUTER SOCIETY CONFERENCE ON COMPUTER VISION AND PATTERN RECOGNITION, 2014, Columbus, Ohio. **Proceedings...** IEEE, 2014. p.580–587.

GONZALEZ, R. C.; WOODS, R. E.; EDDINS, S. L. **Digital Image Processing Using MATLAB**. USA: Prentice-Hall, Inc., 2003.

Gonzalez, R.; Woods, R. **Processamento Digital de Imagens**. 3ª.ed. São Paulo: Pearson Prentice Hall, 2010.

GONZALEZ, S.; ARELLANO, C.; TAPIA, J. E. Deepblueberry: Quantification of Blueberries in the Wild Using Instance Segmentation. **IEEE Access**, Piscataway, New Jersey, v.7, p.105776–105788, 2019.

GOODFELLOW, I.; POUGET-ABADIE, J.; MIRZA, M.; XU, B.; WARDE-FARLEY, D.; OZAIR, S.; COURVILLE, A.; BENGIO, Y. Generative adversarial networks. **Communications of the ACM**, New York, NY, v.63, n.11, p.139–144, 2020.

GU, J.; WANG, Z.; KUEN, J.; MA, L.; SHAHROUDY, A.; SHUAI, B.; LIU, T.; WANG, X.; WANG, L.; WANG, G.; CAI, J.; CHEN, T. **Recent Advances in Convolutional Neural Networks**. Amesterdã: Elsevier, 2017.

GUO, H.; MAO, Y.; ZHANG, R. **MixUp as Locally Linear Out-Of-Manifold Regularization**.

GUO, H.; MAO, Y.; ZHANG, R. **Augmenting Data with Mixup for Sentence Classification**: An Empirical Study.

GUO, S.; WANG, K.; KANG, H.; LIU, T.; GAO, Y.; LI, T. Bin loss for hard exudates segmentation in fundus images. **Neurocomputing**, Amesterdã, v.392, n.xxxx, p.314–324, 2020.

HALEVY, A.; NORVIG, P.; PEREIRA, F. The Unreasonable Effectiveness of Data. **IEEE Intelligent Systems**, Piscataway, New Jersey, v.24, n.2, p.8–12, 2009.

HAO, R.; NAMDAR, K.; LIU, L.; HAIDER, M. A.; KHALVATI, F. **A Comprehensive Study of Data Augmentation Strategies for Prostate Cancer Detection in Diffusion-weighted MRI using Convolutional Neural Networks**.

HAQUE, M. A.; ABDUR RAHMAN, M.; SIDDIK, M. S. Non-Functional Requirements Classification with Feature Extraction and Machine Learning: An Empirical Study. In: INTERNATIONAL CONFERENCE ON ADVANCES IN SCIENCE, ENGINEERING AND ROBOTICS TECHNOLOGY (ICASERT), 1., 2019, Dhaka, Bangladesh. **Proceedings...** IEEE, 2019. p.1–5.

HARNEY, F. Diabetic retinopathy. **Medicine**, Amesterdã, v.34, n.3, p.95–98, 2006.

HASANPOUR, S. H.; ROUHANI, M.; FAYYAZ, M.; SABOKROU, M. **Lets keep it simple, Using simple architectures to outperform deeper and more complex architectures**.

HAWAS, A. R.; ASHOUR, A. S.; GUO, Y. 8 - Neutrosophic set in medical image clustering. In: GUO, Y.; ASHOUR, A. S. (Ed.). **Neutrosophic Set in Medical Image Analysis**. Amesterdã: Elsevier, 2019. p.167–187.

HE, H.; GARCIA, E. A. Learning from Imbalanced Data. **IEEE Transactions on Knowledge and Data Engineering**, Piscataway, New Jersey, v.21, n.9, p.1263–1284, 2009.

HE, K.; GKIOXARI, G.; DOLLÁR, P.; GIRSHICK, R. Mask R-CNN. **IEEE Transactions on Pattern Analysis and Machine Intelligence**, Piscataway, New Jersey, v.42, n.2, p.386–397, 2020.

HE, K.; ZHANG, X.; REN, S.; SUN, J. Spatial Pyramid Pooling in Deep Convolutional Networks for Visual Recognition. **Lecture Notes in Computer Science**, Berlin, p.346–361, 2014.

HE, K.; ZHANG, X.; REN, S.; SUN, J. Deep residual learning for image recognition. In: IEEE COMPUTER SOCIETY CONFERENCE ON COMPUTER VISION AND PATTERN RECOGNITION, 2016, Las Vegas, NV, USA, 27–30 June 2016. **Proceedings...** [S.l.: s.n.], 2016. v.2016-Decem, p.770–778.

HELENE, O.; HELENE, A. F. Alguns aspectos da óptica do olho humano. **Revista Brasileira de Ensino de Fisica**, São Paulo, SP, v.33, n.3, 2011.

HENDRICK, A. M.; GIBSON, M. V.; KULSHRESHTHA, A. Diabetic Retinopathy. **Primary Care - Clinics in Office Practice**, Amesterdã, v.42, n.3, p.451–464, 2015.

HONG, A.; LEE, G.; LEE, H.; SEO, J.; YEO, D. Deep learning model generalization with ensemble in endoscopic images. **CEUR Workshop Proceedings**, Nice, France, v.2886, p.80–89, 2021.

HORRY, M. J.; CHAKRABORTY, S.; PAUL, M.; ULHAQ, A.; PRADHAN, B.; SAHA, M.; SHUKLA, N. COVID-19 Detection Through Transfer Learning Using Multimodal Imaging Data. **IEEE Access**, Piscataway, New Jersey, v.8, p.149808–149824, 2020.

HOSANG, J.; BENENSON, R.; DOLLAR, P.; SCHIELE, B. What Makes for Effective Detection Proposals? **IEEE Transactions on Pattern Analysis and Machine Intelligence**, Piscataway, New Jersey, v.38, n.4, p.814–830, 2016.

HOSSAIN, A.; ISLAM, M. T.; ISLAM, M. S.; CHOWDHURY, M. E. H.; ALMUTAIRI, A. F.; RAZOUQI, Q. A.; MISRAN, N. A YOLOv3 Deep Neural Network Model to Detect Brain Tumor in Portable Electromagnetic Imaging System. **IEEE Access**, Piscataway, New Jersey, v.9, p.82647–82660, 2021.

HSU, W.-Y.; LIN, W.-Y. Adaptive Fusion of Multi-Scale YOLO for Pedestrian Detection. **IEEE Access**, Piscataway, New Jersey, v.9, p.1–1, 2021.

HUANG, G.; LIU, Z.; Van Der Maaten, L.; WEINBERGER, K. Q. Densely connected convolutional networks. In: IEEE CONFERENCE ON COMPUTER VISION AND PATTERN RECOGNITION, CVPR 2017, 30., 2017, Honolulu, USA. **Proceedings...** IEEE, 2017. v.2017-January, p.2261–2269.

HUI, J. **mean Average Precision for Object Detection**. `https://jonathan-hui.medium.com/map-mean-average-precision-for-object-detection-45c121a31173#d9c9`.

HUYNH-THE, T.; LE, B. V.; LEE, S.; LE-TIEN, T.; YOON, Y. **Using weighted dynamic range for histogram equalization to improve the image contrast**. 1–17p. v.2014, n.1.

IACOVACCI, J.; WU, Z.; BIANCONI, G. Mesoscopic structures reveal the network between the layers of multiplex data sets. **Physical Review E - Statistical, Nonlinear, and Soft Matter Physics**, New York, NY, v.92, n.4, p.1–14, 2015.

IBTEHAZ, N.; RAHMAN, M. S. MultiResUNet: Rethinking the U-Net architecture for multimodal biomedical image segmentation. **Neural Networks**, Amesterdã, v.121, p.74–87, 2020.

ICO GUIDELINES FOR DIABETIC EYE CARE. Int. Council Ophthalmol. , Brussels, Belgium, p.1–33, 2017.

ILLINGWORTH, J.; KITTLER, J. The Adaptive Hough Transform. **IEEE Transactions on Pattern Analysis and Machine Intelligence**, Piscataway, New Jersey, v.PAMI-9, n.5, p.690–698, 1987.

IOFFE, S.; SZEGEDY, C. **Batch Normalization**: Accelerating Deep Network Training by Reducing Internal Covariate Shift.

IORJ. **O que é Retina**. [Online; accessed 15-June-2021], https://iorj.med.br/o-que-e-retina/.

IYER, R.; Shashikant Ringe, P.; Varadharajan Iyer, R.; Prabhulal Bhensdadiya, K. Comparison of YOLOv3, YOLOv5s and MobileNet-SSD V2 for Real-Time Mask Detection Comparison of YOLOv3, YOLOv5s and MobileNet-SSD V2 for Real-Time Mask Detection View project Comparison of YOLOv3, YOLOv5s and MobileNet-SSD V2 for Real-Time Mask Detection. **International Journal of Research in Engineering and Technology**, Tamilnadu, n.July, p.1156–1160, 2021.

JADON, S. A survey of loss functions for semantic segmentation. In: IEEE CONFERENCE ON COMPUTATIONAL INTELLIGENCE IN BIOINFORMATICS AND COMPUTATIONAL BIOLOGY, CIBCB 2020, 2020, Virtual Conference. **Proceedings...** IEEE, 2020.

JAKIMOVSKI, G.; DAVCEV, D. Lung cancer medical image recognition using Deep Neural Networks. In: THIRTEENTH INTERNATIONAL CONFERENCE ON DIGITAL INFORMATION MANAGEMENT (ICDIM), 2018, Berlin, Germany. **Proceedings...** IEEE, 2018. p.1–5.

JAPKOWICZ, N. Learning from imbalanced data sets: a comparison of various strategies. **AAAI Workshop on Learning from Imbalanced Data Sets**, Washington, DC, p.0–5, 2000.

JASIM, M. K.; NAJM, R.; KANAN, E. H.; ALFAAR, H. E.; OTAIR, M. **Image Noise Removal Techniques**: A Comparative Analysis. 4–9p. v.8, n.6. Disponível em: <http://www.warse.org/IJSAIT/static/pdf/file/ijsait01862019.pdf>.

JEONG, H.-G.; JEONG, H.-W.; YOON, B.-H.; CHOI, K.-S. Image Segmentation Algorithm for Semantic Segmentation with Sharp Boundaries using Image Processing and Deep Neural Network. In: IEEE INTERNATIONAL CONFERENCE ON CONSUMER ELECTRONICS - ASIA (ICCE-ASIA), 2020, Seoul, Korea South. **Proceedings...** IEEE, 2020. p.1–4.

JIAO, L.; ZHANG, F.; LIU, F.; YANG, S.; LI, L.; FENG, Z.; QU, R. A Survey of Deep Learning-Based Object Detection. **IEEE Access**, Piscataway, New Jersey, v.7, p.128837–128868, 2019.

JOCHER, G. **YOLOv5 releases**. [Online; accessed 04-September-2022], https://github.com/ultralytics/yolov5/releases/.

JONES, J. P.; PALMER, L. A. The two-dimensional spatial structure of simple receptive fields in cat striate cortex. **Journal of Neurophysiology**, Rockville, Maryland, v.58, n.6, p.1187–1211, 1987.

KARKUZHALI, S.; MANIMEGALAI, D. Distinguising Proof of Diabetic Retinopathy Detection by Hybrid Approaches in Two Dimensional Retinal Fundus Images. **Journal of Medical Systems**, Berlin, v.43, n.6, 2019.

KHOJASTEH, P.; Passos Júnior, L. A.; CARVALHO, T.; REZENDE, E.; ALIAHMAD, B.; PAPA, J. P.; KUMAR, D. K. Exudate detection in fundus images using deeply-learnable features. **Computers in Biology and Medicine**, Amesterdã, v.104, p.62–69, 2019.

KIM, J. A.; SUNG, J. Y.; PARK, S. H. Comparison of Faster-RCNN, YOLO, and SSD for Real-Time Vehicle Type Recognition. In: IEEE INTERNATIONAL CONFERENCE ON CONSUMER ELECTRONICS - ASIA, ICCE-ASIA 2020, 2020, Seoul, Korea South. **Proceedings...** IEEE, 2020. p.8–11.

KIM, J.-H.; CHOO, W.; JEONG, H.; SONG, H. O. **Co-Mixup**: Saliency Guided Joint Mixup with Supermodular Diversity.

KINGMA, D. P.; BA, J. **Adam**: A Method for Stochastic Optimization.

KISANTAL, M.; WOJNA, Z.; MURAWSKI, J.; NARUNIEC, J.; CHO, K. **Augmentation for small object detection**.

KONISHI, Y.; HANZAWA, Y.; KAWADE, M.; HASHIMOTO, M. SSD: Single Shot MultiBox Detector. **Eccv**, [S.l.], v.1, p.398–413, 2016.

KRIZHEVSKY, A.; SUTSKEVER, I.; HINTON, G. Imagenet classification with deep convolutional neural networks. **Advances in neural information processing systems**, New York, NY, p.1097–1105, 2012.

KULKARNI, R.; DHAVALIKAR, S.; BANGAR, S. Traffic Light Detection and Recognition for Self Driving Cars Using Deep Learning. In: FOURTH INTERNATIONAL CONFERENCE ON COMPUTING COMMUNICATION CONTROL AND AUTOMATION (ICCUBEA), 2018, Pune, India. **Proceedings...** IEEE, 2018. p.1–4.

KUMAR, D.; ZHANG, X. Improving More Instance Segmentation and Better Object Detection in Remote Sensing Imagery Based on Cascade Mask R-CNN. In: IEEE INTERNATIONAL GEOSCIENCE AND REMOTE SENSING SYMPOSIUM IGARSS, 2021, Brussels, Belgium. **Proceedings...** IEEE, 2021. p.4672–4675.

LABACH, A.; SALEHINEJAD, H.; VALAEE, S. **Survey of Dropout Methods for Deep Neural Networks**.

LAM, T. K.; OHTA, M.; SCHAMONI, S.; RIEZLER, S. **On-the-Fly Aligned Data Augmentation for Sequence-to-Sequence ASR**. New York, NY: Arxiv, 2021. Disponível em: <https://arxiv.org/abs/2104.01393>.

LANDAU, K.; KURZ-LEVIN, M. **Retinal disorders**. 1.ed. Amesterdã: Elsevier, 2011. 97–116p. v.102.

LENTZ, K.; GRIGORYAN, A. A New Measure of Image Enhancement. **IASTED International Conference on Signal Processing & Communication**, [S.l.], p.19–22, 01 2000.

LI, F.-F.; KRISHNA, R.; XU, D. **cs231n, Lecture 15 - Slide 4, Detection and Segmentation**. [Online; accessed 26-December-2021], `http://cs231n.stanford.edu/slides/2021/lecture_15.pdf`.

LI, J.; GUO, S.; KONG, L.; TAN, S.; YUAN, Y. An improved YOLOv3-tiny method for fire detection in the construction industry. In: E3S WEB CONF., 2021, Ternate, Indonesia. **Proceedings...** E3S Web of Conferences, 2021. v.253, p.03069.

LI, L.; CHEN, M.; ZHOU, Y.; WANG, J.; WANG, D. Research of Deep Learning on Gastric Cancer Diagnosis. In: CROSS STRAIT RADIO SCIENCE WIRELESS TECHNOLOGY CONFERENCE (CSRSWTC), 2020, Fuzhou, China. **Proceedings...** IEEE, 2020. p.1–3.

LI, T.; GAO, Y.; WANG, K.; GUO, S.; LIU, H.; KANG, H. Diagnostic assessment of deep learning algorithms for diabetic retinopathy screening. **Information Sciences**, Amesterdã, v.501, p.511–522, 2019.

LI, X.; LAI, T.; WANG, S.; CHEN, Q.; YANG, C.; CHEN, R. Feature Pyramid Networks for Object Detection. In: 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, 2019, Xiamen, China. **Proceedings...** IEEE, 2019. p.1500–1504.

LI, Y. H.; YEH, N. N.; CHEN, S. J.; CHUNG, Y. C. Computer-Assisted Diagnosis for Diabetic Retinopathy Based on Fundus Images Using Deep Convolutional Neural Network. **Mobile Information Systems**, London, UK, v.2019, n.1, 2019.

LI, Y.; QI, H.; DAI, J.; JI, X.; WEI, Y. **Fully Convolutional Instance-aware Semantic Segmentation**.

LI, Z.; YANG, W.; PENG, S.; LIU, F. A Survey of Convolutional Neural Networks: Analysis, Applications, and Prospects. **ArXiv e-prints**, New York, NY, 2020.

LIANG, X.; WU, L.; LI, J.; WANG, Y.; MENG, Q.; QIN, T.; CHEN, W.; ZHANG, M.; LIU, T.-Y. **R-Drop**: Regularized Dropout for Neural Networks.

LIN, K.; ZHAO, H.; LV, J.; ZHAN, J.; LIU, X.; CHEN, R.; LI, C.; HUANG, Z. Face Detection and Segmentation with Generalized Intersection over Union Based on Mask R-CNN. In: ADVANCES IN BRAIN INSPIRED COGNITIVE SYSTEMS: 10TH INTERNATIONAL CONFERENCE, BICS 2019, GUANGZHOU, CHINA, JULY 13–14, 2019„ 2019, Berlin, Heidelberg. **Proceedings...** Springer-Verlag, 2019. p.106–116.

LIN, T. Y.; MAIRE, M.; BELONGIE, S.; HAYS, J.; PERONA, P.; RAMANAN, D.; DOLLÁR, P.; ZITNICK, C. L. Microsoft COCO: Common objects in context. **Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)**, Berlin, v.8693 LNCS, n.PART 5, p.740–755, 2014.

LINDEBERG, T. Scale Invariant Feature Transform. **Scholarpedia**, [S.l.], v.7, n.5, p.10491, 2012.

LIU, C.; JIN, S.; WANG, D.; LUO, Z.; YU, J.; ZHOU, B.; YANG, C. Constrained Oversampling: An Oversampling Approach to Reduce Noise Generation in Imbalanced Datasets with Class Overlapping. **IEEE Access**, Piscataway, New Jersey, p.1–1, 2020.

LIU, S.; QI, L.; QIN, H.; SHI, J.; JIA, J. Path Aggregation Network for Instance Segmentation. In: IEEE COMPUTER SOCIETY CONFERENCE ON COMPUTER VISION AND PATTERN RECOGNITION, 2018, Salt Lake City, USA. **Proceedings...** IEEE, 2018. p.8759–8768.

LIU, Z.; CHEN, W.; ZOU, Y.; HU, C. Regions of interest extraction based on HSV color space. In: IEEE 10TH INTERNATIONAL CONFERENCE ON INDUSTRIAL INFORMATICS, 2012, Beijing, China. **Proceedings...** IEEE, 2012. p.481–485.

LONG, S.; HUANG, X.; CHEN, Z.; PARDHAN, S.; ZHENG, D.; SCALZO, F. Automatic detection of hard exudates in color retinal images using dynamic threshold and SVM classification: Algorithm development and evaluation. **BioMed Research International**, London, UK, v.2019, 2019.

MA, J.; FAN, X.; YANG, S. X.; ZHANG, X.; ZHU, X. Contrast Limited Adaptive Histogram Equalization-Based Fusion in YIQ and HSI Color Spaces for Underwater Image Enhancement. **International Journal of Pattern Recognition and Artificial Intelligence**, Singapura, v.32, n.7, p.1–26, 2018.

MA, J.; YARATS, D. **Quasi-hyperbolic momentum and Adam for deep learning**.

MAIA, N. C. d. F. **Fundamentos básicos da oftalmologia e suas aplicações**. Tocantins: EdUFT, 2018. 122p.

MALOOF, M. Learning When Data Sets are Imbalanced and When Costs are Unequal and Unknown. **Analysis**, Ottawa, v.21, 07 2003.

MAMDOUH, N.; KHATTAB, A. YOLO-Based Deep Learning Framework for Olive Fruit Fly Detection and Counting. **IEEE Access**, Piscataway, New Jersey, v.9, p.84252–84262, 2021.

MANNING, C. D.; RAGHAVAN, P.; SCHüTZE, H. **Introduction to Information Retrieval**. USA: Cambridge University Press, 2008.

MARRONI, L. S. **Aplicação da transformada de Hough para localização dos olhos em faces humanas**. 2002. 103p. Dissertação de Mestrado (Curso de Mestrado em Engenharia Elétrica) — Universidade de São Paulo, São Carlos.

MATEEN, M.; WEN, J.; NASRULLAH, N.; SUN, S.; HAYAT, S. Exudate Detection for Diabetic Retinopathy Using Pretrained Convolutional Neural Networks. **Complexity**, London, UK, v.2020, 2020.

MATUSKA, S.; HUDEC, R.; BENCO, M. The comparison of CPU time consumption for image processing algorithm in Matlab and OpenCV. In: ELEKTRO, 2012, Rajecke Teplice, Slovakia. **Proceedings...** IEEE, 2012. p.75–78.

MCLEOD, D. Why cotton wool spots should not be regarded as retinal nerve fibre layer infarcts. **British Journal of Ophthalmology**, London, UK, v.89, n.2, p.229–237, 2005.

MCREYNOLDS, T.; BLYTHE, D. CHAPTER 12 - Image Processing Techniques. In: MCREYNOLDS, T.; BLYTHE, D. (Ed.). **Advanced Graphics Programming Using OpenGL**. San Francisco: Morgan Kaufmann, 2005. p.211–245. (The Morgan Kaufmann Series in Computer Graphics).

MELO, R.; LIMA, G.; CORRêA, G.; ZATT, B.; AGUIAR, M.; NACHTIGALL, G.; ARAúJO, R. Diagnosis of Apple Fruit Diseases in the Wild with Mask R-CNN. **In: Cerri R., Prati R.C. (eds) Intelligent Systems. BRACIS 2020. Lecture Notes in Computer Science**, Springer: Cham, Switzerland, v.12319 Springer, p.256–270, 2020.

MOHAMMADIAN, S.; KARSAZ, A.; ROSHAN, Y. M. A comparative analysis of classification algorithms in diabetic retinopathy screening. In: INTERNATIONAL CONFERENCE ON COMPUTER AND KNOWLEDGE ENGINEERING (ICCKE), 7., 2017, Mashhad, Iran. **Proceedings...** IEEE, 2017. p.84–89.

MOOKIAH, M. R. K.; ACHARYA, U. R.; CHUA, C. K.; LIM, C. M.; NG, E. Y.; LAUDE, A. Computer-aided diagnosis of diabetic retinopathy: A review. **Computers in Biology and Medicine**, New York, NY, v.43, n.12, p.2136–2155, 2013.

MUKHOPADHYAY, S.; MANDAL, S.; PRATIHER, S.; CHANGDAR, S.; BURMAN, R.; GHOSH, N.; PANIGRAHI, P. K. A comparative study between proposed Hyper Kurtosis based Modified Duo-Histogram Equalization (HKMDHE) and Contrast Limited Adaptive Histogram Equalization (CLAHE) for Contrast Enhancement Purpose of Low Contrast Human Brain CT scan images. **CoRR**, New York, NY, v.abs/1505.06219, 2015.

MURTHY, C. B.; HASHMI, M. F.; BOKDE, N. D.; GEEM, Z. W. Investigations of Object Detection in Images/Videos Using Various Deep Learning Techniques and Embedded Platforms—A Comprehensive Review. **Applied Sciences**, Basel, Switzerland, v.10, n.9, 2020.

NALEPA, J.; MARCINKIEWICZ, M.; KAWULOK, M. Data Augmentation for Brain-Tumor Segmentation: A Review. **Frontiers in Computational Neuroscience**, Lausanne, v.13, p.83, 2019.

NAYAK, J.; BHAT, P. S.; Acharya U, R.; LIM, C. M.; KAGATHI, M. Automated identification of diabetic retinopathy stages using digital fundus images. **Journal of Medical Systems**, Berlin, v.32, n.2, p.107–115, 2008.

NELSON, J. **How to Select the Right Computer Vision Model Architecture**. [Online; accessed 12-June-2021], `https://blog.roboflow.com/yolov3-vs-mobilenet-vs-faster-rcnn/`.

NGUYEN, N. D.; DO, T.; NGO, T. D.; LE, D. D. An Evaluation of Deep Learning Methods for Small Object Detection. **Journal of Electrical and Computer Engineering**, [S.l.], v.2020, 2020.

NGUYEN, T.-S.; STUEKER, S.; NIEHUES, J.; WAIBEL, A. **Improving sequence-to-sequence speech recognition training with on-the-fly data augmentation**. [S.l.]: arXiv, 2019. Disponível em: <https://arxiv.org/abs/1910.13296>.

NIXON, M. S.; AGUADO, A. S. 5 - High-level feature extraction: fixed shape matching. In: NIXON, M. S.; AGUADO, A. S. (Ed.). **Feature Extraction and Image Processing for Computer Vision (Fourth Edition)**. 4.ed. New York, NY: Arxiv, 2020. p.223–290.

NWANKPA, C.; IJOMAH, W.; GACHAGAN, A.; MARSHALL, S. **Activation Functions**: Comparison of trends in Practice and Research for Deep Learning.

OJHA, A.; SAHU, S. P.; DEWANGAN, D. K. Vehicle Detection through Instance Segmentation using Mask R-CNN for Intelligent Vehicle System. In: INTERNATIONAL CONFERENCE ON INTELLIGENT COMPUTING AND CONTROL SYSTEMS (ICICCS), 5., 2021, Madurai, India, 6–8 May 2021. **Proceedings...** IEEE, 2021. p.954–959.

OKSUZ, K.; CAM, B. C.; KAHRAMAN, F.; BALTACI, Z. S.; KALKAN, S.; AKBAS, E. **Mask-aware IoU for Anchor Assignment in Real-time Instance Segmentation**.

OPHTHALMOLOGY, I. C. of. **ICO Guidelines for Diabetic Eye Care,**. [Online; accessed 11-June-2021], http://www.icoph.org/downloads/ICOGuidelinesforDiabeticEyeCare.pdf.

PADILLA, R.; NETTO, S. L.; SILVA, E. A. B. da. A Survey on Performance Metrics for Object-Detection Algorithms. In: INTERNATIONAL CONFERENCE ON SYSTEMS, SIGNALS AND IMAGE PROCESSING (IWSSIP), 2020, Niteroi, Brazil. **Proceedings...** IEEE, 2020. p.237–242.

PAN, S. J.; YANG, Q. A survey on transfer learning. **IEEE Transactions on Knowledge and Data Engineering**, Piscataway, New Jersey, v.22, n.10, p.1345–1359, 2010.

PARIS, S.; DURAND, F. A Fast Approximation of the Bilateral Filter Using a Signal Processing Approach. In: COMPUTER VISION – ECCV 2006, 2006, Berlin, Heidelberg. **Proceedings...** Springer Berlin Heidelberg, 2006. p.568–580.

PARK, G.-H.; CHO, H.-H.; CHOI, M.-R. A contrast enhancement method using dynamic range separate histogram equalization. **IEEE Transactions on Consumer Electronics**, Piscataway, New Jersey, v.54, n.4, p.1981–1987, 2008.

PASZKE, A. et al. **PyTorch**: An Imperative Style, High-Performance Deep Learning Library. Red Hook, NY, USA: Curran Associates Inc., 2019.

PEIXOTO, C. S. B. **Estudo de Métodos de Agrupamento e Transformada de Hough para Processamento de Imagens Digitais**. 2003. 54p. Dissertação de Mestrado (Curso de Mestrado em Matemática) — Universidade Federal da Bahia, Bahia.

PERDOMO, O.; AREVALO, J.; GONZÁLEZ, F. A. Convolutional network to detect exudates in eye fundus images of diabetic subjects. In: INTERNATIONAL SYMPOSIUM ON MEDICAL INFORMATION PROCESSING AND ANALYSIS, 12., 2017, Tandil, Argentina. **Proceedings...** SPIE, 2017. v.10160, p.101600T.

PETERSEN, K.; FELDT, R.; MUJTABA, S.; MATTSSON, M. Systematic Mapping Studies in Software Engineering. In: INTERNATIONAL CONFERENCE ON EVALUATION

AND ASSESSMENT IN SOFTWARE ENGINEERING, 12., 2008, Swindon, GBR. **Proceedings...** BCS Learning & Development Ltd., 2008. p.68–77. (EASE'08).

PHAM, T. Semantic Road Segmentation using Deep Learning. In: APPLYING NEW TECHNOLOGY IN GREEN BUILDINGS (ATIGB), 2021, Da Nang, Vietnam. **Proceedings...** IEEE, 2021. p.45–48.

PHILIP, S.; FLEMING, A. D.; GOATMAN, K. A.; FONSECA, S.; MCNAMEE, P.; SCOTLAND, G. S.; PRESCOTT, G. J.; SHARP, P. F.; OLSON, J. A. The efficacy of automated "disease/no disease"grading for diabetic retinopathy in a systematic screening programme. **British Journal of Ophthalmology**, London, UK, v.91, n.11, p.1512–1517, 2007.

PIRES, R.; ROCHA, A.; WAINER, J. Image Analytics Techniques for Diabetic Retinopathy Detection. In: XXXII CONCURSO DE TESES E DISSERTAçõES, 2019, Porto Alegre, RS, Brasil. **Anais...** SBC, 2019.

PLASTIRAS, G.; KYRKOU, C.; THEOCHARIDES, T. Efficient ConvNet-Based Object Detection for Unmanned Aerial Vehicles by Selective Tile Processing. In: INTERNATIONAL CONFERENCE ON DISTRIBUTED SMART CAMERAS, 12., 2018, New York, NY, USA. **Proceedings...** Association for Computing Machinery, 2018. (ICDSC '18).

POONKASEM, I.; THEERA-UMPON, N.; AUEPHANWIRIYAKUL, S.; PATIKULSILA, D. Detection of hard exudates in fundus images using convolutional neural networks. In: INTERNATIONAL CONFERENCE ON GREEN AND HUMAN INFORMATION TECHNOLOGY, ICGHIT 2019, 7., 2019. **Proceedings...** IEEE, 2019. p.77–81.

PORWAL, P. et al. IDRiD: Diabetic Retinopathy – Segmentation and Grading Challenge. **Medical Image Analysis**, Amesterdã, v.59, 2020.

POTLAPALLY, A.; CHOWDARY, P. S. R.; RAJA SHEKHAR, S.; MISHRA, N.; MADHURI, C. S. V. D.; PRASAD, A. Instance Segmentation in Remote Sensing Imagery using Deep Convolutional Neural Networks. In: INTERNATIONAL CONFERENCE ON CONTEMPORARY COMPUTING AND INFORMATICS (IC3I), 2019, Singapore. **Proceedings...** IEEE, 2019. p.117–120.

POWERS, D. M. W. Evaluation: from precision, recall and F-measure to ROC, informedness, markedness and correlation. **CoRR**, New York, NY, v.abs/2010.16061, 2020.

PRAKASH, N. B.; SELVATHI, D. An efficient approach for detecting exudates in diabetic retinopathy images. **Biomedical Research (India)**, [S.l.], v.2016, n.Special Issue 1, p.S414–S418, 2016.

ORR, G. B.; MÜLLER, K.-R. (Ed.). **Early Stopping - But When?** Berlin, Heidelberg: Springer Berlin Heidelberg, 1998. 55–69p.

PROVOST, F. Machine learning from imbalanced data sets 101. In: AAAI'2000 WORKSHOP ON . . . , 2000, Palo Alto, California. **Proceedings...** AAAI, 2000. p.3.

PUJARI, J.; PUSHPALATHA, S.; PADMASHREE, D. Content-Based Image Retrieval using color and shape descriptors. In: INTERNATIONAL CONFERENCE ON SIGNAL AND IMAGE PROCESSING, 2010, Chennai, India. **Proceedings...** IEEE, 2010. p.239–242.

PUTTEMANS, S.; CALLEMEIN, T.; GOEDEMé, T. Building Robust Industrial Applicable Object Detection Models using Transfer Learning and Single Pass Deep Learning Architectures. In: INTERNATIONAL JOINT CONFERENCE ON COMPUTER VISION, IMAGING AND COMPUTER GRAPHICS THEORY AND APPLICATIONS, 13., 2018. **Proceedings...** SCITEPRESS - Science and Technology Publications, 2018.

QI, D.; TAN, W.; YAO, Q.; LIU, J. YOLO5Face: Why Reinventing a Face Detector. **ArXiv e-prints**, New York, NY, 2021.

QUMMAR, S.; KHAN, F. G.; SHAH, S.; KHAN, A.; SHAMSHIRBAND, S.; REHMAN, Z. U.; KHAN, I. A.; JADOON, W. A Deep Learning Ensemble Approach for Diabetic Retinopathy Detection. **IEEE Access**, Piscataway, New Jersey, v.7, p.150530–150539, 2019.

RAHMAN, R.; AZAD, Z. B.; HASAN, M. B. Densely-Populated Traffic Detection using YOLOv5 and Non-Maximum Suppression Ensembling. **ArXiv e-prints**, New York, NY, 2021.

RAI, R.; GOUR, P.; SINGH, B. Underwater Image Segmentation using CLAHE Enhancement and Thresholding. **International Journal of Emerging Technology and Advanced Engineering**, [S.l.], v.2, p.118–123, 01 2012.

RAMCHARAN, A.; MCCLOSKEY, P.; BARANOWSKI, K.; MBILINYI, N.; MRISHO, L.; NDALAHWA, M.; LEGG, J.; HUGHES, D. Assessing a mobile-based deep learning model for plant disease surveillance. **CoRR**, New York, NY, v.abs/1805.08692, 2018.

REDMON, J. **Darknet**: Open Source Neural Networks in C. `http://pjreddie.com/darknet/`.

REDMON, J.; DIVVALA, S.; GIRSHICK, R.; FARHADI, A. You only look once: Unified, real-time object detection. In: IEEE COMPUTER SOCIETY CONFERENCE ON COMPUTER VISION AND PATTERN RECOGNITION, 2016, Las Vegas, NV, USA, 27–30 June 2016. **Proceedings...** IEEE, 2016. v.2016-Decem, p.779–788.

REDMON, J.; FARHADI, A. YOLO9000: Better, Faster, Stronger. In: IEEE CONFERENCE ON COMPUTER VISION AND PATTERN RECOGNITION (CVPR), 2017, Honolulu, USA. **Proceedings...** IEEE, 2017. p.6517–6525.

REDMON, J.; FARHADI, A. YOLOv3: An Incremental Improvement. **ArXiv e-prints**, New York, NY, 2018. arXiv:1804.02767.

REN, S.; HE, K.; GIRSHICK, R.; SUN, J. Faster R-CNN: Towards Real-Time Object Detection with Region Proposal Networks. **IEEE Transactions on Pattern Analysis and Machine Intelligence**, Piscataway, New Jersey, v.39, n.6, p.1137–1149, 2017.

REYES, A. K.; CAICEDO, J. C.; CAMARGO, J. E. Fine-tuning Deep Convolutional Networks for Plant Recognition. In: CONFERENCE AND LABS OF THE EVALUATION FORUM, WORKING NOTES OF CLEF, TOULOUSE, FRANCE, SEPTEMBER 8-11, 2015, 2015. **Proceedings...** CEUR-WS.org, 2015. (CEUR Workshop Proceedings, v.1391).

REZATOFIGHI, H.; TSOI, N.; GWAK, J.; SADEGHIAN, A.; REID, I.; SAVARESE, S. Generalized intersection over union: A metric and a loss for bounding box regression. In: IEEE COMPUTER SOCIETY CONFERENCE ON COMPUTER VISION AND PATTERN RECOGNITION, 2019, Long Beach, USA. **Proceedings...** IEEE, 2019. v.2019-June, p.658–666.

RIBEIRO, A. M.; JUNIOR, F. d. P. S. A. Um Estudo Comparativo Entre Cinco Métodos de Otimização Aplicados Em Uma RNC Voltada ao Diagnóstico do Glaucoma. **Revista de Sistemas e Computação - RSC**, Salvador, v.10, n.1, p.122–130, 2020.

RICHARDS, E.; MUNAKOMI, S.; MATHEW, D. **Optic Nerve Sheath Ultrasound.** Online; accessed 15-Dec-2023, `https://www.ncbi.nlm.nih.gov/books/NBK554479/`.

RIORDAN-EVA, P.; AUGSBURGER, J. J. **General Ophthalmology**. 19.ed. New York, NY, USA: Mc Graw Hill Education, 2018.

ROBERTS, D. A.; YAIDA, S.; HANIN, B. **The Principles of Deep Learning Theory**.

RONG, F.; DU-WU, C.; BO, H. A Novel Hough Transform Algorithm for Multi-objective Detection. In: THIRD INTERNATIONAL SYMPOSIUM ON INTELLIGENT INFORMATION TECHNOLOGY APPLICATION, 2009, Nanchang, China. **Proceedings...** IEEE, 2009. v.3, p.705–708.

RONNEBERGER, O.; FISCHER, P.; BROX, T. U-net: Convolutional networks for biomedical image segmentation. **Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)**, Berlin, v.9351, p.234–241, 2015.

RUDER, S. An overview of gradient descent optimization algorithms. **ArXiv e-prints**, New York, NY, p.1–14, 2016. arXiv:1609.04747.

RUDER, S. **An overview of gradient descent optimization algorithms**.

S, P. T.; MAI, F.; VOGELS, T.; JAGGI, M.; FLEURET, F. **On the Tunability of Optimizers in Deep Learning**. Disponível em: <https://openreview.net/forum?id=H1gEP6NFwr>.

SABATINI, S. P. Recurrent inhibition and clustered connectivity as a basis for Gabor-like receptive fields in the visual cortex. **Biological Cybernetics**, Berlin, v.74, n.3, p.189–202, 1996.

SANTOS, C.; AGUIAR, M.; WELFER, D.; BELLONI, B. A New Approach for Detecting Fundus Lesions Using Image Processing and Deep Neural Network Architecture Based on YOLO Model. **Sensors**, Basel, Switzerland, v.22, n.17, 2022.

SANTOS, C.; DE AGUIAR, M. S.; WELFER, D.; BELLONI, B. Deep Neural Network Model based on One-Stage Detector for Identifying Fundus Lesions. In: INTERNATIONAL JOINT CONFERENCE ON NEURAL NETWORKS (IJCNN), 2021, Shenzhen, China, 18–22 July 2021. **Proceedings...** IEEE, 2021. p.1–8.

SANTOS, J. R. V. dos. **Avaliação de técnicas de realce de imagens digitais utilizando métricas subjetivas e objetivas**. 2016. 82p. Dissertação de Mestrado (Engenharia de Teleinformática) — Universidade Federal do Ceará, Fortaleza.

SANTOS, V. A. **Curso de Deep Learning - Object Detection- SSD Fast Faster RCNN Yolo**. [Online; accessed 12-June-2021], `https://ww2.inf.ufg.br/~anderson/deeplearning/20181/`.

SCHETTINI, R.; GASPARINI, F.; CORCHS, S.; MARINI, F.; CAPRA, A.; CASTORINA, A. Contrast image correction method. **J. Electronic Imaging**, Miami, Florida, v.19, p.023005, 04 2010.

SETIAWAN, A. W.; MENGKO, T. R.; SANTOSO, O. S.; SUKSMONO, A. B. Color retinal image enhancement using CLAHE. In: INTERNATIONAL CONFERENCE ON ICT FOR SMART SOCIETY 2013: "THINK ECOSYSTEM ACT CONVERGENCE", ICISS 2013, 2013, Jakarta, Indonesia. **Proceedings...** IEEE, 2013. p.215–217.

SHAH, V.; KYRILLIDIS, A.; SANGHAVI, S. **Minimum weight norm models do not always generalize well for over-parameterized problems**.

SHAO, L.; ZHU, F.; LI, X. Transfer learning for visual categorization: A survey. **IEEE Transactions on Neural Networks and Learning Systems**, Piscataway, New Jersey, v.26, n.5, p.1019–1034, 2015.

SHARIF, M.; AMIN, J.; YASMIN, M.; REHMAN, A. Efficient hybrid approach to segment and classify exudates for DR prediction. **Multimedia Tools and Applications**, Berlin, v.79, n.15-16, p.11107–11123, 2020.

SHELHAMER, E.; LONG, J.; DARRELL, T. Fully Convolutional Networks for Semantic Segmentation. **IEEE Transactions on Pattern Analysis and Machine Intelligence**, Piscataway, New Jersey, v.39, n.4, p.640–651, 2017.

SHENAVARMASOULEH, F.; MOHAMMADI, F. G.; AMINI, M. H.; TAHA, T.; RASHEED, K.; ARABNIA, H. R. **DRDrV3**: Complete Lesion Detection in Fundus Images Using Mask R-CNN, Transfer Learning, and LSTM. [S.l.]: arXiv, 2021. Disponível em: <https://arxiv.org/abs/2108.08095>.

SHENE, C.-K. **Geometric Transformations**. Online; accessed 01-Nov-2021, `https://pages.mtu.edu/~shene/COURSES/cs3621/NOTES/geometry/geo-tran.html`.

SHIAO, Y. H.; CHEN, T. J.; CHUANG, K. S.; LIN, C. H.; CHUANG, C. C. Quality of compressed medical images. **Journal of Digital Imaging**, Berlin, v.20, n.2, p.149–159, 2007.

SHILOH-PERL, L.; GIRYES, R. **Introduction to deep learning**.

SHORTEN, C.; KHOSHGOFTAAR, T. M. A survey on Image Data Augmentation for Deep Learning. **Journal of Big Data**, Berlin, v.6, n.1, 2019.

SILVA, A. B. **Métodos Computacionais para Análise e Classificação de Displasias em Imagens da Cavidade Bucal**. 2019. 102p. Dissertação de Mestrado (Curso de Ciência da Computação) — Universidade Federal de Uberlândia, Uberlândia.

SIMONYAN, K.; ZISSERMAN, A. Very deep convolutional networks for large-scale image recognition. In: INTERNATIONAL CONFERENCE ON LEARNING REPRESENTATIONS, ICLR 2015 - CONFERENCE TRACK PROCEEDINGS, 3., 2015, San Diego, CA, USA, 7–9 May 2015. **Proceedings...** ICLR, 2015. p.1–14.

SINGH, P. K.; TIWARI, V. Normalized Log Twicing Function for DC Coefficients Scaling in LAB Color Space. In: INTERNATIONAL CONFERENCE ON INVENTIVE RESEARCH IN COMPUTING APPLICATIONS, ICIRCA 2018, 2018. **Proceedings...** IEEE, 2018. n.Icirca, p.333–338.

SOLAWETZ, J. **YOLOv5: The Latest Model for Object Detection**. [Online; accessed 31-May-2021], `https://blog.roboflow.com/yolov5-improvements-and-evaluation/`.

SON, J.; SHIN, J. Y.; KIM, H. D.; JUNG, K. H.; PARK, K. H.; PARK, S. J. Development and Validation of Deep Learning Models for Screening Multiple Abnormal Findings in Retinal Fundus Images. **Ophthalmology**, Amesterdã, v.127, n.1, p.85–94, 2020.

SRIVASTAVA, N.; HINTON, G.; KRIZHEVSKY, A.; SUTSKEVER, I.; SALAKHUTDINOV, R. Dropout: A Simple Way to Prevent Neural Networks from Overfitting. **Journal of Machine Learning Research**, New York, NY, v.15, n.56, p.1929–1958, 2014.

SUN, C.; SHRIVASTAVA, A.; SINGH, S.; GUPTA, A. Revisiting Unreasonable Effectiveness of Data in Deep Learning Era. In: IEEE INTERNATIONAL CONFERENCE ON COMPUTER VISION, 2017, Venice, Italy. **Proceedings...** IEEE, 2017. v.2017-October, p.843–852.

SUN, K.; WANG, B.; ZHOU, Z.-Q.; ZHENG, Z.-H. Real time image haze removal using bilateral filter. **Transactions of Beijing Institute of Technology**, [S.l.], v.31, p.810–814, 07 2011.

SUTSKEVER, I.; MARTENS, J.; DAHL, G.; HINTON, G. On the importance of initialization and momentum in deep learning. In: INTERNATIONAL CONFERENCE ON MACHINE LEARNING, 30., 2013, Atlanta, Georgia, USA. **Proceedings...** PMLR, 2013. n.3, p.1139–1147. (Proceedings of Machine Learning Research, v.28).

SZEGEDY, C.; VANHOUCKE, V.; IOFFE, S.; SHLENS, J.; WOJNA, Z. Rethinking the Inception Architecture for Computer Vision. In: IEEE COMPUTER SOCIETY CONFERENCE ON COMPUTER VISION AND PATTERN RECOGNITION, 2016, Las Vegas, USA. **Proceedings...** IEEE, 2016. v.2016-December, p.2818–2826.

TAN, L.; JIANG, J. **Image Processing Basics**. Amesterdã: Elsevier, 2019. 649–726p.

TAN, M.; PANG, R.; LE, Q. V. EfficientDet: Scalable and efficient object detection. In: IEEE COMPUTER SOCIETY CONFERENCE ON COMPUTER VISION AND PATTERN RECOGNITION, 2020, Seattle, USA. **Proceedings...** IEEE, 2020. p.10778–10787.

TAQI, A. M.; AWAD, A.; AL-AZZO, F.; MILANOVA, M. The Impact of Multi-Optimizers and Data Augmentation on TensorFlow Convolutional Neural Network Performance. In: IEEE CONFERENCE ON MULTIMEDIA INFORMATION PROCESSING AND RETRIEVAL (MIPR), 2018, Miami, USA. **Proceedings...** IEEE, 2018. p.140–145.

THEERA-UMPON, N.; POONKASEM, I.; AUEPHANWIRIYAKUL, S.; PATIKULSILA, D. Hard exudate detection in retinal fundus images using supervised learning. **Neural Computing and Applications**, Berlin, v.32, n.17, p.13079–13096, 2020.

TIAN, J.; YUAN, J.; LIU, H. Road Marking Detection Based on Mask R-CNN Instance Segmentation Model. In: INTERNATIONAL CONFERENCE ON COMPUTER VISION, IMAGE AND DEEP LEARNING (CVIDL), 2020, Chongqing, China. **Proceedings...** IEEE, 2020. p.246–249.

TING, D. S. W.; CHEUNG, G. C. M.; WONG, T. Y. Diabetic retinopathy: global prevalence, major risk factors, screening practices and public health challenges: a review. **Clinical and Experimental Ophthalmology**, Hoboken, New Jersey, v.44, n.4, p.260–277, 2016.

ULLAH, H.; SABA, T.; ISLAM, N.; ABBAS, N.; REHMAN, A.; MEHMOOD, Z.; ANJUM, A. An ensemble classification of exudates in color fundus images using an evolutionary algorithm based optimal features selection. **Microscopy Research and Technique**, Hoboken, New Jersey, v.82, n.4, p.361–372, 2019.

UNEL, F. O.; OZKALAYCI, B. O.; CIGLA, C. The Power of Tiling for Small Object Detection. In: IEEE/CVF CONFERENCE ON COMPUTER VISION AND PATTERN RECOGNITION WORKSHOPS (CVPRW), 2019, Long Beach, USA. **Proceedings...** IEEE, 2019. p.582–591.

VOCATURO, E.; ZUMPANO, E. The contribution of AI in the detection of the Diabetic Retinopathy. In: IEEE INTERNATIONAL CONFERENCE ON BIOINFORMATICS AND BIOMEDICINE, BIBM 2020, 2020, Seoul, Korea, 16–19 December 2020. **Proceedings...** IEEE, 2020. p.1516–1519.

VRBANCIC, G.; PODGORELEC, V. Transfer Learning With Adaptive Fine-Tuning. **IEEE Access**, Piscataway, New Jersey, v.8, p.196197–196211, 2020.

WALTER, T.; KLEIN, J. C.; MASSIN, P.; ERGINAY, A. A contribution of image processing to the diagnosis of diabetic retinopathy - Detection of exudates in color fundus images of the human retina. **IEEE Transactions on Medical Imaging**, Piscataway, New Jersey, v.21, n.10, p.1236–1243, 2002.

WANG, C. Y.; Mark Liao, H. Y.; WU, Y. H.; CHEN, P. Y.; HSIEH, J. W.; YEH, I. H. CSPNet: A new backbone that can enhance learning capability of CNN. In: IEEE COMPUTER SOCIETY CONFERENCE ON COMPUTER VISION AND PATTERN RECOGNITION WORKSHOPS, 2020, Virtual Conference. **Proceedings...** IEEE, 2020. v.2020-June, p.1571–1580.

WANG, H.; YUAN, G.; ZHAO, X.; PENG, L.; WANG, Z.; HE, Y.; QU, C.; PENG, Z. Hard exudate detection based on deep model learned information and multi-feature joint representation for diabetic retinopathy screening. **Computer Methods and Programs in Biomedicine**, Amesterdã, v.191, p.105398, 2020.

WANG, K.; FANG, B.; QIAN, J.; YANG, S.; ZHOU, X.; ZHOU, J. Perspective Transformation Data Augmentation for Object Detection. **IEEE Access**, Piscataway, New Jersey, v.8, p.4935–4943, 2020.

WANG, S.; ZHENG, J.; HU, H.-M.; LI, B. Naturalness Preserved Enhancement Algorithm for Non-Uniform Illumination Images. **IEEE Transactions on Image Processing**, Piscataway, New Jersey, v.22, n.9, p.3538–3548, 2013.

WANGENHEIM, A. **Avaliando, Validando e Testando o seu Modelo: Metodologias de Avaliação de Performance**. [Online; accessed 13-June-2021], `https://bit.ly/3ES5c0D`.

WARNER, R. Measurement of Meat Quality | Measurements of Water-holding Capacity and Color: Objective and Subjective. In: DIKEMAN, M.; DEVINE, C. (Ed.). **Encyclopedia of Meat Sciences (Second Edition)**. 2.ed. Oxford: Academic Press, 2014. p.164–171.

WEISS, G. M. Mining with Rarity: A Unifying Framework. **SIGKDD Explor. Newsl.**, New York, NY, USA, v.6, n.1, p.7–19, June 2004.

WEISS, K.; KHOSHGOFTAAR, T. M.; WANG, D. D. **A survey of transfer learning**. [S.l.]: Springer International Publishing, 2016. v.3, n.1.

WILLIAMS, R.; AIREY, M.; BAXTER, H.; FORRESTER, J.; KENNEDY-MARTIN, T.; GIRACH, A. Epidemiology of diabetic retinopathy and macular oedema: A systematic review. **Eye**, Berlin, v.18, n.10, p.963–983, 2004.

WILSON, A. C.; ROELOFS, R.; STERN, M.; SREBRO, N.; RECHT, B. **The Marginal Value of Adaptive Gradient Methods in Machine Learning**.

WU, Y.; KIRILLOV, A.; MASSA, F.; LO, W.-Y.; GIRSHICK, R. **Detectron2**. `https://github.com/facebookresearch/detectron2`.

XIAO, Y.; WATSON, M. Guidance on Conducting a Systematic Literature Review. **Journal of Planning Education and Research**, [S.l.], v.39, n.1, p.93–112, 2019.

XIE, J.; ZHENG, S. **ZSD-YOLO**: Zero-Shot YOLO Detection using Vision-Language KnowledgeDistillation.

XIE, S.; GIRSHICK, R.; DOLLÁR, P.; TU, Z.; HE, K. Aggregated residual transformations for deep neural networks. In: IEEE CONFERENCE ON COMPUTER VISION AND PATTERN RECOGNITION, CVPR 2017, 30., 2017, Honolulu, USA. **Proceedings...** IEEE, 2017. v.2017-January, p.5987–5995.

XIE, S.; TU, Z. Holistically-nested edge detection. **Proceedings of the IEEE International Conference on Computer Vision**, Santiago, Chile, 7–13 December 2015, v.2015 Inter, p.1395–1403, 2015.

XU, K.; FENG, D.; MI, H. Deep convolutional neural network-based early automated detection of diabetic retinopathy using fundus image. **Molecules**, Basel, Switzerland, v.22, n.12, 2017.

XU, R.; LIN, H.; LU, K.; CAO, L.; LIU, Y. A forest fire detection system based on ensemble learning. **Forests**, Basel, Switzerland, v.12, n.2, p.1–17, 2021.

YADAV, G.; MAHESHWARI, S.; AGARWAL, A. Contrast limited adaptive histogram equalization based enhancement for real time video system. In: INTERNATIONAL CONFERENCE ON ADVANCES IN COMPUTING, COMMUNICATIONS AND INFORMATICS (ICACCI), 2014, Delhi, India. **Proceedings...** IEEE, 2014. p.2392–2397.

YANG, Q.; TAN, K.-H.; AHUJA, N. Real-time O(1) bilateral filtering. In: IEEE CONFERENCE ON COMPUTER VISION AND PATTERN RECOGNITION, 2009, Miami, USA. **Proceedings...** IEEE, 2009. p.557–564.

YE, H.; SHANG, G.; WANG, L.; ZHENG, M. A new method based on hough transform for quick line and circle detection. In: INTERNATIONAL CONFERENCE ON BIOMEDICAL ENGINEERING AND INFORMATICS (BMEI), 8., 2015, Shenyang, China. **Proceedings...** IEEE, 2015. p.52–56.

YE, L.; ZHU, W.; FENG, S.; CHEN, X. GANet: Group attention network for diabetic retinopathy image segmentation. In: MEDICAL IMAGING 2020: IMAGE PROCESSING, 2020, Houston, USA. **Proceedings...** SPIE, 2020. v.1131307, n.March, p.5.

YE, Z.; MOHAMADIAN, H.; YE, Y. Discrete Entropy and Relative Entropy Study on Nonlinear Clustering of Underwater and Arial Images. In: IEEE INTERNATIONAL CONFERENCE ON CONTROL APPLICATIONS, 2007, Singapore. **Proceedings...** IEEE, 2007. p.313–318.

YEN, G. G.; LEONG, W. F. A sorting system for hierarchical grading of diabetic fundus images: A preliminary study. **IEEE Transactions on Information Technology in Biomedicine**, [S.l.], v.12, n.1, p.118–130, 2008.

YOSINSKI, J.; CLUNE, J.; BENGIO, Y.; LIPSON, H. How transferable are features in deep neural networks? **Advances in Neural Information Processing Systems**, Piscataway, New Jersey, v.4, n.January, p.3320–3328, 2014.

YU, Y.; ZHAO, J.; GONG, Q.; HUANG, C.; ZHENG, G.; MA, J. Real-Time Underwater Maritime Object Detection in Side-Scan Sonar Images Based on Transformer-YOLOv5. **Remote Sensing**, Piscataway, New Jersey, v.13, n.18, 2021.

YUDITA, S. I.; MANTORO, T.; AYU, M. A. Deep Face Recognition for Imperfect Human Face Images on Social Media using the CNN Method. In: INTERNATIONAL CONFERENCE OF COMPUTER AND INFORMATICS ENGINEERING (IC2IE), 4., 2021, Depok, Indonesia. **Proceedings...** IEEE, 2021. p.412–417.

YUEN, H.; PRINCEN, J.; ILLINGWORTH, J.; KITTLER, J. Comparative study of Hough Transform methods for circle finding. **Image and Vision Computing**, Amesterdã, v.8, n.1, p.71–77, 1990.

ZAGORUYKO, S.; KOMODAKIS, N. **Wide Residual Networks**.

ZAMIR, A.; SAX, A.; SHEN, W.; GUIBAS, L.; MALIK, J.; SAVARESE, S. Taskonomy: Disentangling task transfer learning. In: INTERNATIONAL JOINT CONFERENCE ON ARTIFICIAL INTELLIGENCE, 2019, Macao. **Proceedings...** IJCAI, 2019. v.2019-August, p.6241–6245.

ZEILER, M. D.; FERGUS, R. Visualizing and Understanding Convolutional Networks. In: EUROPEAN CONFERENCE ON COMPUTER VISION – ECCV 2014, 2014, Cham. **Proceedings...** Springer International Publishing, 2014. p.818–833.

ZEILER, M. D.; TAYLOR, G. W.; FERGUS, R. Adaptive deconvolutional networks for mid and high level feature learning. In: INTERNATIONAL CONFERENCE ON COMPUTER VISION, 2011, Barcelona, Spain. **Proceedings...** IEEE, 2011. p.2018–2025.

ZHANG, C.; BENGIO, S.; HARDT, M.; RECHT, B.; VINYALS, O. **Understanding deep learning requires rethinking generalization**.

ZHANG, H.; CISSE, M.; DAUPHIN, Y. N.; LOPEZ-PAZ, D. MixUp: Beyond empirical risk minimization. In: INTERNATIONAL CONFERENCE ON LEARNING REPRESENTATIONS, ICLR 2018 - CONFERENCE TRACK PROCEEDINGS, 6., 2018, Vancouver, Canada. **Proceedings...** ICLR, 2018. p.1–13.

ZHANG, Q.; ZHANG, M.; CHEN, T.; SUN, Z.; MA, Y.; YU, B. Recent advances in convolutional neural network acceleration. **Neurocomputing**, [S.l.], v.323, p.37–51, 2019.

ZHANG, X.; GWEON, H.; PROVOST, S. Threshold Moving Approaches for Addressing the Class Imbalance Problem and their Application to Multi-label Classification. **PervasiveHealth: Pervasive Computing Technologies for Healthcare**, New York, NY, v.PartF169255, p.72–77, 2020.

ZHAO, H.; LI, Q.; FENG, H. Multi-Focus Color Image Fusion in the HSI Space Using the Sum-Modified-Laplacian and a Coarse Edge Map. **Image Vision Comput.**, USA, v.26, n.9, p.1285–1295, Sept. 2008.

ZHAO, Z.-Q.; ZHENG, P.; XU, S. tao; WU, X. **Object Detection with Deep Learning**: A Review.

ZHENG, Z.; WANG, P.; LIU, W.; LI, J.; YE, R.; REN, D. Distance-IoU loss: Faster and better learning for bounding box regression. In: AAAI CONFERENCE ON ARTIFICIAL INTELLIGENCE (AAAI), 2020, Palo Alto, USA. **Proceedings...** AAAI, 2020.

ZHENG, Z.; ZHAO, J.; LI, Y. Research on Detecting Bearing-Cover Defects Based on Improved YOLOv3. **IEEE Access**, Piscataway, New Jersey, v.9, p.10304–10315, 2021.

ZHOU, Z.-H.; LIU, X.-Y. Training cost-sensitive neural networks with methods addressing the class imbalance problem. **IEEE Transactions on Knowledge and Data Engineering**, Piscataway, New Jersey, v.18, n.1, p.63–77, 2006.

ZHU, L.; GENG, X.; LI, Z.; LIU, C. Improving YOLOv5 with Attention Mechanism for Detecting Boulders from Planetary Images. **Remote Sensing**, Basel, Switzerland, v.13, n.18, 2021.

ZHU, X.; LYU, S.; WANG, X.; ZHAO, Q. **TPH-YOLOv5**: Improved YOLOv5 Based on Transformer Prediction Head for Object Detection on Drone-captured Scenarios.

ZIMMERMANN, R. S.; SIEMS, J. N. Faster training of Mask R-CNN by focusing on instance boundaries. **Computer Vision and Image Understanding**, Amesterdã, v.188, p.102795, 2019.

ZOU, L. Chapter 5 - Meta-learning for computer vision. In: ZOU, L. (Ed.). **Meta-Learning**. [S.l.]: Academic Press, 2023. p.91–208.

ZOU, Z.; SHI, Z.; GUO, Y.; YE, J. **Object Detection in 20 Years**: A Survey.

Anexos

ANEXO A – CLASSIFICAÇÃO DA RETINOPATIA DIABÉTICA DE ACORDO COM A PRESENÇA DE CARACTERÍSTICAS CLÍNICAS

De acordo com Alghadyan (2011); Etdrsr (1991a); Philip et al. (2007); Etdrsr (1991b), os tipos e estágios da RD podem ser classificados conforme as características clínicas apresentadas na Tabela 36.

Tabela 36 – Estágios da Retinopatia Diabética: Retinopatia Diabética Não-Proliferativa (RDNP), Retinopatia Diabética Proliferativa (RDP) e Edema Macular Diabético (EMD).

Tipos de RD	**Estágios**	**Lesões de Fundo**
RDNP	Leve (Veja a Figura 60b)	- MA, HE, EX e EMD.
	Moderada (Veja a Figura 60c)	- Difusão de HE e/ou MA e/ou SE. - VB ou IRMA.
	Severa (Veja a Figura 60d)	- MA, HE, SE, VB presentes em pelo menos dois quadrantes da retina. - HE em quatro quadrantes. - VB em dois quadrantes. - IRMA grave em um quadrante.
RDP (Veja a Figura 60e)	RDP inicial	- HE pré-retinianas.
	Alto risco de RDP	- HE Vítreas. - NV do DO. - NV em qualquer lugar no DO.
	RDP avançada	- Descolamento da Retina. - NV da Íris.
EMD (Veja a Figura 60f)	EMD clinicamente não significativo	- Presença de edema, espessamento da retina ou EX a mais de 500 mícrons da fóvea.
	EMD clinicamente significativo	- Presença de edema, espessamento da retina ou EX na fóvea ou a pelo menos 500 mícrons desta.

A Figura 60 apresenta as imagens de fundo: (a) Normal, (b) RDNP leve, (c) RDNP moderada, (d) RDNP grave, (e) RDP e (f) Edema Macular Diabético (MOOKIAH et al., 2013).

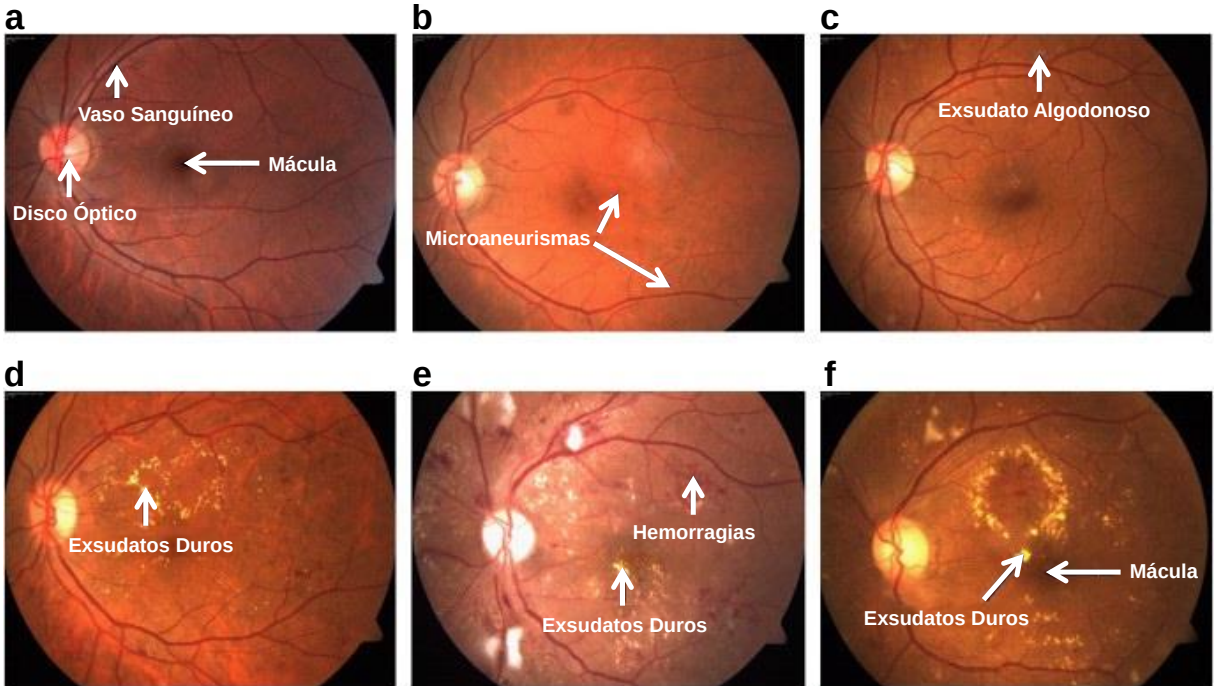


Figura 60 – Imagens de fundo: (a) Normal; (b) RDNP leve; (c) RDNP moderada; (d) RDNP grave; (e) RDP; (f) Edema Macular Diabético. Fonte: Adaptado de Mookiah et al. (2013).

ANEXO B – CARACTERÍSTICAS DA RETINOPATIA DIABÉTICA - *INTERNATIONAL COUNCIL OF OPHTHALMOLOGY*

O Conselho Internacional de Oftalmologia (em inglês, *Internacional Council of Ophthalmology* - ICO) (*ICO GUIDELINES FOR DIABETIC EYE CARE*, 2017), desenvolveu as Diretrizes para o Tratamento do Olho Diabético para desempenhar um papel educacional e de apoio para oftalmologistas e prestadores de cuidados com os olhos mundialmente. Essas diretrizes têm a intenção de melhorar a qualidade dos cuidados com os olhos de pacientes ao redor do mundo.

As Diretrizes foram concebidas para informar os oftalmologistas sobre as exigências para triagem e detecção de Retinopatia Diabética e a avaliação e gerenciamento apropriados de pacientes com RD. As Diretrizes demonstram, também, a necessidade de os oftalmologistas trabalharem com prestadores de cuidados primários, bem como com especialistas adequados, como endocrinologistas.

Com o problema de Diabetes e a Retinopatia Diabética crescendo rapidamente no mundo, é vital assegurar que oftalmologistas e provedores de cuidados com os olhos estejam adequadamente preparados.

As Diretrizes foram criadas para ser um documento de trabalho e serão atualizadas constantemente. Elas foram lançadas primeiramente em dezembro de 2013 e atualizado em janeiro de 2017.

De acordo com as diretrizes do ICO (*ICO GUIDELINES FOR DIABETIC EYE CARE*, 2017), as características da Retinopatia Diabética são apresentadas na Tabela 37.

Tabela 37 – Características da Retinopatia Diabética (*ICO GUIDELINES FOR DIABETIC EYE CARE*, 2017).

Característica	Descrição	Considerações de Avaliação
Microaneurismas	Pontos vermelhos, esféricos e isolados, de tamanhos variados. Podem refletir uma tentativa abortada de formação de um novo vaso ou podem simplesmente ser uma fraqueza da parede do vaso capilar por meio da perda da integridade estrutural normal.	São os mais fáceis de serem observados em angiografia fluorescente.
Hemorragias puntiformes	Hemorragias de ponto nem sempre podem ser diferenciadas de microaneurismas por serem similares em aparência, mas com tamanho variado.	O termo Hemorragia de Ponto/Microaneurisma (H/Ma) é normalmente usado.
Hemorragias de mancha	Formadas onde aglomerados de capilares obstruem, levando à formação de hemorragias de mancha intrarretinianas.	A lesão pode parecer estar na camada plexiforme externa em angiografia fluorescente, onde ela não mascara o leito capilar sobrejacente, ao contrário de hemorragias de ponto e de chama, que se encontram mais superficialmente na retina.
Pontos algodonosos	Representam as extremidades inchadas de axônios interrompidos, onde ocorre o acúmulo de fluxo axoplasmático na borda do infarto.	Essas características não são exclusivas de RD e não parecem aumentar, por si mesmas, o risco de formação de novos vasos. Por exemplo, podem ocorrer em hipertensão de HIV/AIDS.
Anomalias intrarretinianas microvasculares	São remanescente capilares dilatados seguindo fechamentos extensos de redes capilares entre arteríola e vênula. Características relacionadas incluem: • perolização venosa (foco de proliferação de célula endotelial venosa que falhou em se desenvolver em novos vasos), • Reduplicação venosa (rara), • Circuitos venosos (pensado para se desenvolver devido a pequena oclusão de vaso e abertura de circulação alternativa) e, • Palidez da retina e vasos brancos.	São os mais fáceis de serem observados em angiografia fluorescente.
Alterações maculares em Retinopatia não proliferativa – Edema macular – Doença macrovascular	Espessamento da retina acontece devido à acumulação de fluidos de exsudatos originados na barreira sangue-retina exterior danificada (edema extracelular) ou como resultado de hipoxia, levando à acumulação de fluido dentro de células retinianas individuais (edema intracelular). Pode ser focal ou difuso. Hemorragia de chama e formação de ponto algodonoso. Pode ocorrer devido a oclusão arteriolar, sem oclusão capilar, o que frequentemente afeta a camada horizontal de fibras nervosas da retina.	A aparência de edema macular pode ser observada em exame estereoscópico ou inferida pela presença de exsudato intrarretiniano.
Alterações no disco óptico	Discos ópticos inchados podem, ocasionalmente, ser vistos (papilopatia diabética) em pacientes diabéticos.	Em papilopatia diabética, a visão, geralmente, não é significativamente prejudicada.

ANEXO C – CARACTERÍSTICAS DA RETINOPATIA DIABÉTICA PROLIFERATIVA - *INTERNATIONAL COUNCIL OF OPHTHALMOLOGY*

De acordo com as diretrizes do ICO (*ICO GUIDELINES FOR DIABETIC EYE CARE*, 2017), as características da Retinopatia Diabética proliferativa são apresentadas na Tabela 38.

Tabela 38 – Características da Retinopatia Diabética Proliferativa (*ICO GUIDELINES FOR DIABETIC EYE CARE*, 2017).

Característica	**Descrição**	**Considerações de Avaliação**
Neovasos no disco (NVD)	Neovasos no disco geralmente surgem da circulação venosa no disco ou a 1 diâmetro de disco do disco.	Para diferenciar NVD de pequenos vasos sanguíneos bem normais, note que os últimos sempre afinam para uma extremidade não voltam ao início do disco, enquanto NVD sempre volta ao início, podendo formar uma rede caótica na dentro da volta, e tem a parte superior do laço com diâmetro mais largo que a base.
Neovasos em outros lugares (NVOL)	Neovasos, que geralmente ocorrem ao longo da fronteira entre a retina saudável e áreas de oclusão capilar.	Não devem ser confundidos com anormalidades microvasculares intrarretinianas, as quais ocorrem dentro de áreas de oclusão capilar.
Outros locais de neovasos	Formação de neovasos na íris (NVI) é incomum, mas representa potencialmente mais alterações isquêmicas avançadas. Formação de novos vasos na superfície hialóide anterior ocorre, raramente, após vitrectomia, se quantidade de laser insuficiente for aplicada na retina periférica.	É útil executar gonioscopia em tais casos, para excluir novos vasos no ângulo da câmara anterior (NVA), os quais podem levar ao glaucoma neovascular.
Proliferação fibrosa	Em Retinopatia Proliferativa, novos vasos crescem sobre uma plataforma de células gliais.	