Universidade de São Paulo
Instituto de Matemática e Estatística
Bacharelado em Ciência da Computação

Reconhecimento automático de retinopatia diabética via imagem de fundo de olho

Aya Meira, Lucas Sobrinho, Miqueias Lima

Monografia Final

mac 499 — Trabalho de Formatura Supervisionado

Supervisora: Prof.ª Dr.ª Nina S. T. Hirata

2024

Resumo

Aya Meira, Lucas Sobrinho, Miqueias Lima. **Reconhecimento automático de retinopatia diabética via imagem de fundo de olho**. Monografia (Bacharelado). Instituto de Matemática e Estatística, Universidade de São Paulo, São Paulo, 2024.

A retinopatia diabética é uma das principais causas de cegueira em todo o mundo. A detecção precoce é fundamental para prevenir danos visuais graves. Este trabalho explora o uso de modelos de aprendizado de máquina para o reconhecimento automatizado da retinopatia diabética, destacando sua eficácia, limitações e áreas de melhoria. Por meio de uma revisão de métodos de última geração e experimentos em conjuntos de dados públicos, este estudo visa a contribuir para o avanço do uso da inteligência artificial no apoio à decisão clínica em oftalmologia.

Palavras-chave: Retinopatia Diabética. Aprendizado Profundo. Visão Computacional. Redes Neurais Convolucionais. Classificação de Imagens. Oftalmologia.

Abstract

Aya Meira, Lucas Sobrinho, Miqueias Lima. **Automatic detection of diabetic retinopathy via fundus imaging**. Capstone Project Report (Bachelor). Institute of Mathematics and Statistics, University of São Paulo, São Paulo, 2024.

Diabetic retinopathy is one of the leading causes of blindness worldwide. Early detection is critical to preventing serious visual damage. This work explores the use of machine learning models for automated recognition of diabetic retinopathy, highlighting their effectiveness, limitations and areas for improvement. Through a review of state-of-the-art methods and experiments on public datasets, this study aims to contribute to advancing the use of artificial intelligence in clinical decision support in ophthalmology.

Keywords: Diabetic Retinopathy. Deep Learning. Computer Vision. Convolutional Neural Networks. Image Classification. Ophthalmology.

Lista de figuras

Lista de tabelas

Sumário

Capítulo 1

Introdução

1.1 Definição do Problema

A retinopatia diabética (RD) afeta milhões de pessoas globalmente e é uma complicação séria da diabetes. O diagnóstico tradicional da doença é feito por inspeção visual de imagens de fundo de olho, buscando indícios de lesões na retina. Este método depende da avaliação especializada por oftalmologistas, o que pode ser demorado e de difícil acesso, especialmente em regiões em que o acesso a serviços médicos é mais escasso. Apesar da alta prevalência da diabetes, doença a qual a RD está relacionada, estratégias de tratamento são efetivas em 90% dos casos para prevenir perda visual severa, especialmente se for detectada precocemente (Ophthalmology AAO, 2016).

Redes Neurais Convolucionais (CNNs, do inglês Convolutional Neural Networks) destacam-se pelo seu desempenho em tarefas que envolvem o processamento de imagens, como classificação, segmentação e detecção de padrões complexos em dados estruturados em grade (LeCun *et al.*, 1998; Krizhevsky *et al.*, 2012). Essas características tornam as CNNs especialmente promissoras na análise de exames médicos de imagem, como radiografias, tomografias computadorizadas, ressonâncias magnéticas e, no caso deste trabalho, imagens de fundo de olho. Este trabalho investiga o uso de modelos de Deep Learning (DL) para automatizar a detecção e classificação dos estágios da retinopatia diabética (RD).

A aplicação de CNNs na área médica tem demonstrado grande potencial em apoiar profissionais de saúde na tomada de decisão, reduzindo a carga de trabalho e aumentando a precisão diagnóstica (Li *et al.*, 2019; Galdran *et al.*, 2021). Em muitos casos, as CNNs alcançam níveis de desempenho comparáveis ou superiores aos especialistas humanos, especialmente quando treinadas com grandes volumes de dados anotados (Teo *et al.*, 2021; Martinez-Murcia *et al.*, 2021). No campo da oftalmologia, o uso de CNNs para o diagnóstico e classificação de glaucoma e retinopatia diabética já foi amplamente investigado, com resultados promissores em diferentes estudos (Porwa *et al.*, 2020; Sebastian *et al.*, 2023; L. F. Nakayama *et al.*, 2023).

O diagnóstico da retinopatia diabética é realizado por inspeção visual da imagem de fundo do olho, avaliando a presença de certos tipos de anomalias na retina. A presença e distribuição de cada tipo de anomalia determina o grau da doença conforme o International

Clinical Diabetic Retinopathy (ICDR), um protocolo clínico bem estabelecido na medicina (Ophthalmology AAO, 2016).

Idealmente, a RD deve ser monitorada em pacientes com diabetes para possibilitar intervenções precoces. Devido ao método de inspeção visual, esse monitoramento requer a presença de oftalmologistas ou profissionais treinados. Quando consideradas situações de grande escala de pacientes, tais como atendimento em postos de saúde públicos, em geral, não se pode contar com a presença contínua e em suficiente número destes profissionais.

Diante deste cenário, há grande interesse no desenvolvimento de sistemas computacionais para auxílio ao diagnóstico da RD. De fato, já existem soluções na indústria que realizam a detecção e avaliação automática desta patologia aprovadas pelo Food and Drug Administration (FDA) (Grzybowski e Brona, 2023) e usadas comercialmente nos EUA e Europa.

A área médica também apresenta desafios específicos. Uma característica desejável em sistemas de auxílio a diagnóstico é que eles sejam capazes também de apontar evidências que justifiquem o resultado. Isto é, é muito mais significativo quando um sistema computacional indica que há grande possibilidade de que exista uma anomalia de um certo tipo e mostre, simultaneamente, onde tal anomalia se encontra na imagem. As evidências apontadas poderão servir para melhor fundamentar o diagnóstico médico.

Para desenvolver sistemas com esse tipo de capacidade, a máquina precisa ser treinada para detectar as anomalias. Isto, por sua vez, requer que as imagens sejam anotadas não apenas quanto ao tipo de lesões presentes, mas também quanto à localização de cada uma das lesões. Dentre as formas para anotar a localização, os mais comuns são os retângulos envoltórios (bounding boxes) e a segmentação semântica (ao nível de pixel). O segundo tipo de anotação, mais refinada, é importante se existe necessidade de quantificar a extensão (em termos de área afetada) das anomalias.

1.2 Visão geral do trabalho

O objetivo deste trabalho foi desenvolver um modelo de aprendizado de máquina capaz de realizar o reconhecimento de retinopatia diabética a partir da imagem de fundo de olho.

Para isso, treinamos modelos de deep learning utilizando três arquiteturas distintas, avaliando seu desempenho na detecção de retinopatia diabética. Além disso, validamos técnicas para lidar com o desbalanceamento dos dados e propusemos um novo método para mitigar esse problema. Como resultado, obtivemos um modelo capaz de alcançar 97,5% de acurácia e 89,7% de F1-score, demonstrando a eficácia das abordagens adotadas.

O presente texto está organizado em 3 partes: Investigação Teórica, Metodologia e Experimentos. Na primeira parte foram investigados técnicos e teóricos relevantes da retinopatia diabética (capítulo 2) e desenvolvida a fundamentação teórica sobre técnicas de aprendizado de máquina (capítulo 3). Na segunda parte, foram analisados trabalhos relacionados (capítulo 4), examinando recursos e abordagens comumente usados na literatura e descrição das metodologias usadas (capítulo 5). Por fim, são apresentados os experimentos conduzidos, bem como os resultados, discussões e dificuldades encontrados (capítulo 6).

Capítulo 2

Retinopatia Diabética

A retinopatia diabética é uma complicação microvascular crônica da diabetes que afeta os vasos sanguíneos da retina, que é a camada sensível à luz no fundo do olho. Essa condição é causada por níveis elevados de glicose no sangue, que, ao longo do tempo, danificam as paredes dos capilares retinianos. Esses danos resultam em vazamentos de fluidos, hemorragias e, em estágios avançados, crescimento anormal de vasos sanguíneos, levando a possíveis complicações graves, como descolamento de retina, edema macular diabético (EMD) e cegueira (Bowling, 2020).

A diabetes provoca alterações metabólicas que levam ao estresse oxidativo e inflamação na retina, resultando na disfunção endotelial. Os capilares tornam-se mais permeáveis, levando à formação de microaneurismas e hemorragias. Em estágios avançados, ocorre neovascularização, onde novos vasos sanguíneos anormais crescem, mas são frágeis e mais propensos a romper.

A RD é geralmente dividida em duas categorias principais: Retinopatia Diabética Não Proliferativa (RDNP) e Retinopatia Diabética Proliferativa (RDP), que incluem, respectivamente, as formas menos graves e as mais graves de RD.

A bibliografia médica produziu diferentes formas de classificar a RD. Uma delas, usada neste trabalho, foi desenvolvida pelo Congresso Internacional de Oftalmologia em 2012, sendo atualmente denominada *International Clinical Diabetic Retinopathy (ICDR) severity scale* (American Academy of Ophthalmology, 2021). A escala foi desenvolvida com o propósito de simplificar estudos anteriores, de tal forma que fosse fácil de adotar a escala na prática clínica. De acordo com este padrão, a gravidade da DR pode ser classificada em cinco categorias (Hernández, Smith *et al.*, 2023).

2.0.1 Tipos de lesões na retinopatia diabética

A retinopatia diabética é caracterizada por uma variedade de lesões que refletem o dano microvascular na retina. A tarefa de aprendizado de máquina sobre a qual este trabalho se debruça é treinar um modelo de visão computacional capaz de identificar automaticamente a presença e o grau da retinopatia diabética, a qual é diagnosticada a partir da observação destas lesões em imagens de fundo de olho. A título de ilustração, as principais formas

de lesões observadas incluem:

Microaneurismas

Os microaneurismas são as primeiras alterações visíveis na retinopatia diabética. Eles representam dilatações localizadas nos capilares retinianos, causadas pela fraqueza da parede dos vasos. Esses pontos aparecem como manchas vermelhas circulares em exames de fundo de olho, como ilustrado na figura 2.1.

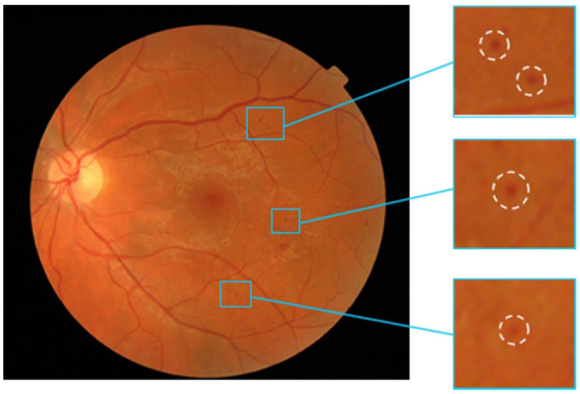


Figura 2.1: *Microaneurismas destacados em imagem de fundo de olho.*
Fonte: *Wu et al., 2017*

Hemorragias Retinianas

As hemorragias retinianas ocorrem devido à ruptura de capilares frágeis. Podem ser classificadas em:

- **Hemorragias em chama:** Lesões alongadas que se espalham nas camadas mais superficiais da retina, em direção às fibras nervosas (ver figura 2.2).
- **Hemorragias em ponto ou mancha:** Lesões pequenas e circulares geralmente localizadas nas camadas internas da retina (ver figura 2.3).

Figura 2.2: *Hemorragia retiniana em chama em imagem de fundo de olho.*
Fonte: *Neimark, 2024*

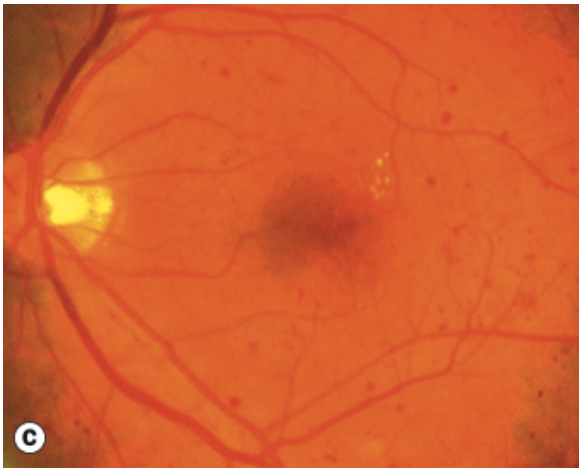


Figura 2.3: *Hemorragia retiniana em ponto ou mancha.*
Fonte: *(Bowling, 2020, P. 524)*

Exsudatos duros

Os exsudatos duros são depósitos lipídicos e proteicos que se acumulam na retina como consequência do vazamento dos capilares danificados. Eles aparecem como manchas amarelas brilhantes (ver figura 2.4), frequentemente associadas ao edema macular diabético (EMD).

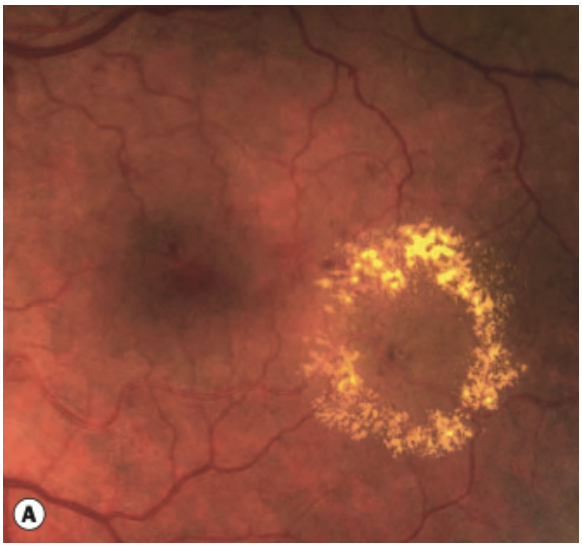


Figura 2.4: *Anel de exsudatos duros.*
Fonte: *Bowling, 2020, P. 526*

Manchas algodonosas (exsudatos moles)

As manchas algodonosas são áreas esbranquiçadas e opacas causadas pela isquemia localizada e pelo bloqueio do fluxo axoplasmático. São sinais de isquemia grave e geralmente indicam um estágio mais avançado da RD (ver figura 2.5).

Figura 2.5: *Exsudato mole.*
Fonte: *ASRS, s.d.*

Neovascularização

A neovascularização é característica da retinopatia diabética proliferativa (RDP). Trata-se do crescimento de novos vasos sanguíneos frágeis e anormais como resposta à isquemia, como nas figuras 2.6 e 2.7. Esses vasos podem se romper facilmente, resultando em hemorragia vítrea ou pré-retiniana.

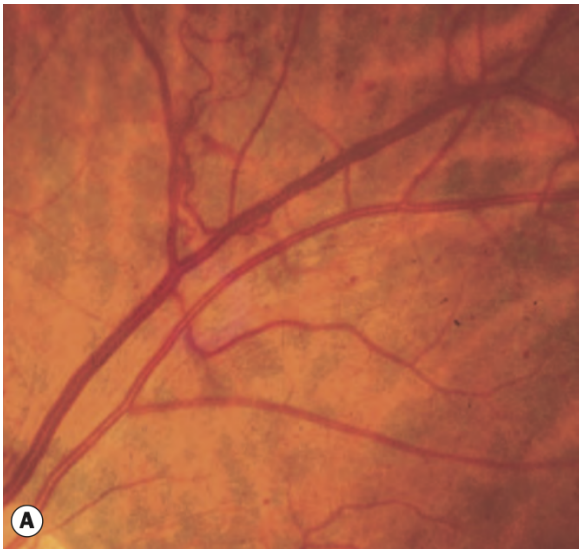


Figura 2.6: *Neovascularização suave.*
Fonte: *Bowling, 2020, P. 526*

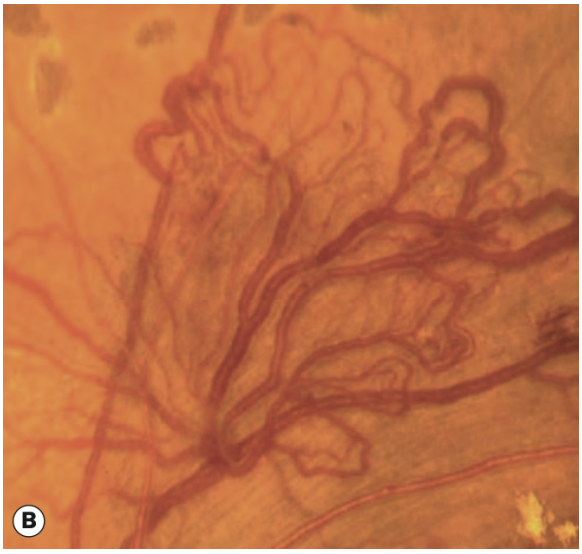


Figura 2.7: *Neovascularização severa.*
Fonte: *Bowling, 2020, P. 526*

Edema Macular Diabético (EMD)

O EMD é causado pelo acúmulo de fluido na mácula, região central da retina responsável pela visão detalhada. Ele é um dos principais responsáveis pela perda de visão em pacientes com retinopatia diabética. Aparece frequentemente em pacientes com RDNP moderada a grave e pode ocorrer mesmo sem sinais de retinopatia proliferativa (ver figura 2.8).

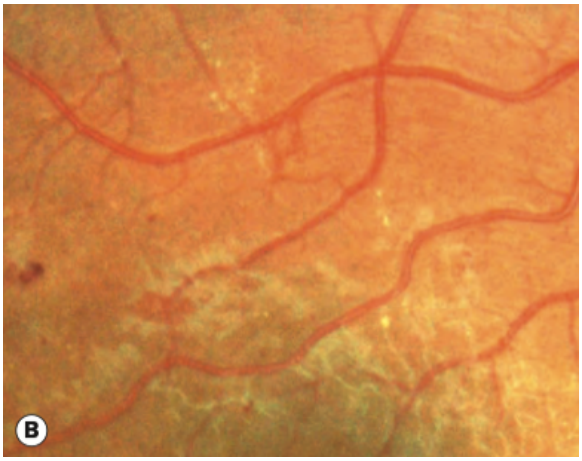


Figura 2.8: *Aparência clínica do edema.*
Fonte: *Bowling, 2020, P. 529*

IRMA (Anomalias Microvasculares Intrarretinianas)

As IRMAs representam canais microvasculares que se desenvolvem dentro da retina para compensar a isquemia. São sinais de retinopatia não proliferativa grave e podem preceder o desenvolvimento de neovascularização.

Hemorragia Vítrea

A hemorragia vítrea ocorre quando os vasos sanguíneos anormais da neovascularização se rompem e sangram no vítreo, o gel que preenche o interior do olho (ver figura 2.9). É uma complicação séria que pode causar perda de visão súbita.

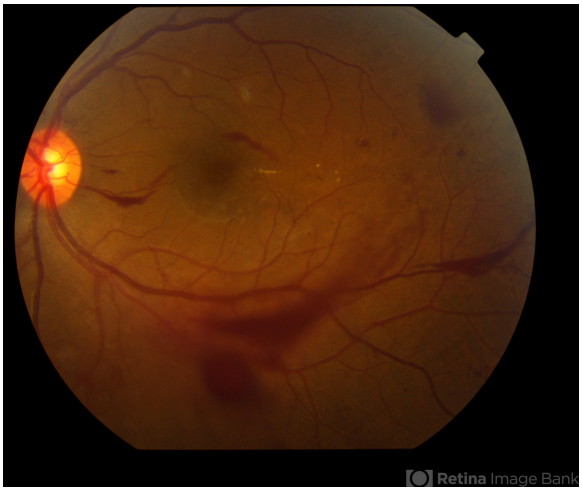


Figura 2.9: *Hemorragia vítrea em imagem de fundo de olho de paciente de retinopatia diabética.* ***Fonte:*** *ASRS, s.d.*

A identificação das lesões indicadas acima leva à classificação da RD nas categorias da tabela 2.1:

Categoria	Descrição
0. Sem DR aparente	Nenhum sinal visível de retinopatia diabética.
1. DR leve	Apenas microaneurismas ou hemorragias com ou sem exsudatos duros.
2. DR moderada	4 ou mais hemorragias do tipo "ponto"ou "mancha"em apenas 1 hemicampo[a]
3. DR severa	4 ou mais hemorragias nos hemicampos inferior e superior.
4. DR proliferativa	Presença de neovascularização e/ou hemorragia vítrea ou pré-retiniana.

Tabela 2.1: *Classificação da gravidade da Retinopatia Diabética segundo o ICDR.*

[a] Hemicampo é uma metade do campo de visão.

2.1 Impacto e prevalência

Estima-se que, globalmente, aproximadamente 103 milhões de pessoas vivam com RD, das quais cerca de 28 milhões apresentam formas avançadas, como a retinopatia diabética proliferativa (RDP) ou o edema macular diabético (EMD). Esses dados foram apresentados

por Teo et al. em 2021, que realizaram uma análise abrangente sobre a prevalência global da RD (Teo *et al.*, 2021).

O aumento desses números está diretamente relacionado ao crescimento da prevalência da diabetes mellitus. De acordo com a International Diabetes Federation (IDF), em 2021, havia aproximadamente 537 milhões de pessoas vivendo com diabetes em todo o mundo. A IDF projeta que esse número pode alcançar 783 milhões até 2045, evidenciando uma tendência preocupante no aumento de casos de diabetes e, consequentemente, de suas complicações, como a RD (International Diabetes Federation, 2021).

No Brasil, uma revisão sistemática com meta-análise estimou que aproximadamente 36,28% dos adultos brasileiros com diabetes apresentam algum grau de retinopatia diabética (RD). De acordo com a International Diabetes Federation (IDF), cerca de 15,7 milhões de pessoas no Brasil vivem com diabetes (Chagas *et al.*, 2023). Com base nesses dados, pode-se estimar que aproximadamente 5,65 milhões de brasileiros convivam com algum nível de retinopatia diabética.

Capítulo 3

Fundamentação teórica das redes neurais convolucionais

3.1 Introdução

Nos últimos anos, as redes neurais convolucionais se estabeleceram como uma das arquiteturas mais poderosas e versáteis para o processamento de dados estruturados em forma de grade, tais como imagens, sinais unidimensionais e até mesmo séries temporais. Desde sua introdução e popularização na década de 1990, com os trabalhos pioneiros de LeCun e colaboradores (LeCun *et al.*, 1998), as CNNs passaram por diversas transformações, evoluindo em complexidade, capacidade de generalização e eficiência computacional. A principal razão para sua proeminência reside na capacidade de extrair automaticamente características hierárquicas de alto nível a partir dos dados brutos, reduzindo a dependência de engenharia manual de atributos (feature engineering) e tornando-se o motor de muitas aplicações bem-sucedidas em visão computacional, reconhecimento de fala, processamento de linguagem natural e outras áreas correlatas.

A partir de 2012, com o marco histórico da rede AlexNet no desafio ImageNet Large Scale Visual Recognition Challenge (ILSVRC) Krizhevsky *et al.*, 2012, as CNNs obtiveram uma visibilidade global no contexto acadêmico e industrial. De lá para cá, arquiteturas mais profundas e complexas, como VGG, Inception, ResNet e DenseNet, demonstraram sucessivos avanços em tarefas de classificação, detecção e segmentação de imagens, além de aplicações em reconhecimento facial, análise médica de imagens, veículos autônomos e uma infinidade de problemas que se beneficiam da extração de padrões visuais.

Este capítulo tem como objetivo apresentar os principais conceitos teóricos, matemáticos e arquiteturais que sustentam as redes neurais convolucionais. Partiremos de uma introdução ao conceito de convolução, suas propriedades e o porquê de essa operação ser tão valiosa no contexto do aprendizado profundo. Em seguida, discutiremos o papel de camadas essenciais, como camadas convolucionais, de pooling e de ativação, bem como seus fundamentos matemáticos. Por fim, revisaremos algumas das arquiteturas clássicas que, historicamente, moldaram o campo e inspiraram o design das redes modernas.

3.2 Principais conceitos e fundamentos matemáticos

A essência das redes neurais convolucionais repousa sobre o operador matemático conhecido como convolução. Na análise de sinais e processamento digital de imagens, a convolução é uma operação linear que descreve como a forma de um sinal de saída resulta da combinação entre um sinal de entrada e um filtro (ou núcleo ou kernel) de convolução. No contexto das CNNs, a convolução é utilizada para detectar padrões locais nos dados, reduzindo a dependência da posição exata desses padrões no espaço (translational invariance).

3.2.1 Operação de convolução

A notação para a convolução de f e g é $(f * g)$. Ela é definida como a integral do produto de uma das funções por uma cópia deslocada e invertida da outra:

$$s(t) = (f * g)(t) := \int_{-\infty}^{\infty} f(\tau)g(t-\tau)d\tau$$

Para criar a intuição da convolução, digamos que tenhamos as funções $f(\tau)$ e $g(t)$, onde $f(\tau)$ representa uma medida no momento τ e $g(t)$ representa um peso atribuído a esta medida em relação a um intervalo de tempo t (entendido como a idade da medida). Podemos interpretar a convolução como uma forma de gerar uma estimativa suavizada de $f(\tau)$ para um dado instante t, ponderando a contribuição de $f(\tau)$ de acordo com sua "idade"em relação ao momento atual t, conforme determinado pela função $g(t-\tau)$.

Em abordagens mais simples de aprendizado de máquina, muitas vezes não nos preocupamos com a ordem dos elementos nos vetores de dados usados como entrada para o treinamento. No entanto, ao lidar com imagens, um pixel está fortemente correlacionado com seus pixels vizinhos. Essas correlações locais são importantes. As CNNs aproveitam esse fato para reduzir o número de parâmetros necessários e oferecer maior precisão na generalização.

O valor da função de convolução, definida por uma integral, pode ser interpretado como a "área de correspondência sob a curva"entre o sinal original e o kernel (filtro). Intuitivamente, essa interseção está relacionada à similaridade entre o sinal e o filtro.

Do ponto de vista matemático, se considerarmos uma imagem I de duas dimensões (por exemplo, uma matriz $I(i,j)$ de dimensões $H \times W$ onde H é a altura e W a largura), e um filtro F de dimensões $k \times k$, a operação de convolução bidimensional pode ser expressa como:

$$G(i,j) = (I * F)(i,j) = \sum_{m=0}^{k-1} \sum_{n=0}^{k-1} I(i+m, j+n)F(m,n),$$

onde $G(i,j)$ é a saída da convolução no ponto (i,j). Na prática, as convoluções utilizadas em CNNs costumam ser restritas a filtros relativamente pequenos (como 3×3, 5×5, etc.) e são aplicadas a cada canal da imagem de entrada, combinando-os para produzir mapas

de ativação. A ideia central é que cada filtro aprende um conjunto específico de características do dado de entrada, por exemplo, bordas horizontais, verticais ou texturas mais complexas. Conforme se avança nas camadas, esses filtros tornam-se progressivamente mais sofisticados, representando características de alto nível.

A convolução possui uma série de propriedades matemáticas importantes. Uma delas é a linearidade, o que significa que o resultado da convolução de uma soma de sinais com um dado filtro é igual à soma das convoluções individuais. Além disso, a convolução é comutativa, ou seja, $I * F = F * I$. No entanto, no contexto das CNNs, normalmente pensamos na operação de convolução como um deslizamento do filtro sobre a imagem, o que facilita o raciocínio sobre a localização espacial das características extraídas.

3.2.2 Camadas convolucionais e pesos treináveis

Uma camada convolucional em uma CNN consiste em um conjunto de filtros (kernels) cujo valores são parâmetros ajustáveis, aprendidos durante o treinamento da rede. Diferentemente de transformações lineares densas (como as transformações presentes nas camadas totalmente conectadas), em que cada entrada está ligada a cada saída, a convolução impõe uma conexão local: cada unidade de saída está conectada apenas a uma pequena região da entrada, determinada pelo tamanho do kernel. Essa restrição local ajuda a capturar padrões espaciais e a reduzir dramaticamente o número de parâmetros, um fator crucial para viabilizar o treinamento de redes profundas.

Se, por exemplo, um filtro F de tamanho $k \times k$ for aplicado a uma imagem com três canais de cor (RGB), então o filtro terá dimensão $k \times k \times 3$. Assim, o número de parâmetros para um filtro é $k^2 \times C$, onde C é o número de canais da entrada. Se a camada convolucional possui M filtros, o número total de parâmetros será $M \times (k^2 \times C)$. Esse número é tipicamente bem menor do que o de uma camada densa correspondente, pois não há conexão entre todos os pixels e todas as unidades de saída, apenas entre regiões locais.

3.2.3 Funções de ativação não-lineares

Após a operação linear de convolução, a saída passa por uma função de ativação não-linear. Essa etapa é essencial para injetar não-linearidade na rede neural, de tal forma que ela seja capaz de aproximar funções complexas e não apenas transformações lineares. As funções de ativação mais populares em CNNs são a ReLU (Rectified Linear Unit), a qual é definida como:

$$\mathrm{ReLU}(x) = \max(0, x).$$

A ReLU possui diversas vantagens em relação a funções mais antigas, como a sigmoide e a tanh: ela não sofre com o problema de gradientes muito pequenos (vanishing gradients (Glorot e Bengio, 2010)) na mesma intensidade e acelera o treinamento (Nair e Hinton, 2010). Outras variantes como Leaky ReLU, ELU e GELU também são utilizadas, dependendo do cenário. Essas funções tornam as relações entre entradas e saídas não-lineares, aumentando o poder expressivo da rede.

3.2.4 Pooling e redução dimensional

Outra operação fundamental nas CNNs é o *pooling*. O pooling é um processo de subamostragem espacial que reduz a dimensão da representação, retendo informação estatística relevante e proporcionando invariância a pequenas translações. As formas mais comuns são o *max pooling*, que seleciona o valor máximo em uma região local, e o *average pooling*, que calcula a média dos valores. Por exemplo, se considerarmos um *max pooling* com janela 2×2, aplicado a um mapa de características, a saída terá metade da largura e metade da altura do mapa original, resultando em uma redução significativa de parâmetros e custo computacional.

Ao reduzir gradualmente a dimensão espacial, as camadas de pooling permitem que as camadas posteriores da rede foquem em características de mais alto nível, sem escalar o custo computacional. Além disso, a operação de pooling ajuda a tornar a rede mais robusta a pequenas variações na posição das características detectadas, aumentando a capacidade de generalização (LeCun *et al.*, 1998).

3.2.5 Regularização, normalização e inicialização

Além de convolução, ativação e pooling, existem outras técnicas importantes no projeto de CNNs. Por exemplo, a regularização é essencial para evitar sobreajuste (overfitting) (Goodfellow *et al.*, 2016). Métodos como a desativação parcial de unidades (*dropout*) reduzem a coadaptação entre neurônios, melhorando a capacidade de generalização (Srivastava *et al.*, 2014). Outra técnica relevante é a normalização por lotes (*batch normalization*), que reduz o deslocamento interno da distribuição dos dados e acelera o treinamento, estabilizando o fluxo de gradientes (Ioffe e Szegedy, 2015).

A inicialização dos pesos dos filtros também desempenha um papel crítico. Inicializações mal planejadas podem levar a gradientes explosivos ou a gradientes que se anulam rapidamente, dificultando o treinamento. Métodos como a inicialização de He, Glorot (Xavier) e outras heurísticas têm sido propostos para garantir que a variação do sinal seja preservada ao longo das camadas, tornando o treinamento mais estável.

3.2.6 Aprendizado através do Gradiente Descendente

O treinamento de CNNs, assim como outras redes neurais, é fundamentado no gradiente descendente. O erro é medido por uma função de custo, comumente a entropia cruzada no caso de classificação, e a atualização dos pesos é feita via retropropagação (backpropagation). Essa técnica envolve aplicar a regra da cadeia para computar a derivada do erro em relação a cada peso, propagando os gradientes das camadas de saída para as camadas iniciais (Goodfellow *et al.*, 2016).

Há muitas variantes do gradiente descendente, tais como o SGD (Stochastic Gradient Descent) com momento (Polyak, 1964), Adam (Kingma e Ba, 2015), Adagrad (Duchi *et al.*, 2011), entre outras, cada uma com suas próprias características e hiperparâmetros que influenciam a velocidade de convergência e a qualidade do ponto ótimo encontrado.

3.2.7 Equivariança por translação e compartilhamento de parâmetros

Um dos aspectos mais significativos das CNNs é sua capacidade de detectar padrões independentemente de sua localização na imagem (Goodfellow *et al.*, 2016). Isso é alcançado graças a duas propriedades chaves: a operação de convolução, que é local e translacionalmente invariante, e o compartilhamento de parâmetros, no qual o mesmo filtro é aplicado em todas as posições da entrada. Enquanto as camadas totalmente conectadas teriam parâmetros distintos para cada posição do pixel, uma camada convolucional utiliza o mesmo conjunto de pesos (kernel) para todas as posições. Isso significa que a rede pode "reconhecer" a mesma característica em qualquer lugar do campo de visão, resultando em maior eficiência e menor número de parâmetros.

Este compartilhamento de parâmetros e a invariância espacial tornam as CNNs particularmente adequadas para imagens, em que um mesmo padrão (por exemplo, uma aresta ou um canto) pode aparecer em múltiplas regiões. É esse atributo fundamental que diferencia as CNNs dos modelos densos, tornando-as mais escaláveis para entradas de dimensões maiores.

3.3 Técnicas avançadas

3.3.1 Balanced MixUp

Esse mecanismo proposto em Galdran *et al.*, 2021 tem por objetivo mitigar o sub aproveitamento de modelos de aprendizado de máquina em tarefas de classificação em datasets desbalanceadas (classes com cardinalidade muito diferentes). Ela é baseada no popular método de aumento MixUp, mas propõe processos de amostragens diferentes.

MixUp

O MixUp Zhang *et al.*, 2018 gera uma nova instância X_c a partir da interpolação convexa entre dois elementos X_a e X_b amostradas aleatoriamente de um conjunto de dados $\mathcal{D}$. Considere α fixo e λ amostrado da distribuição Beta, temos:

$$X_c = \lambda \cdot X_a + (1 - \lambda) \cdot X_b; \text{ com } \lambda \sim Beta(\alpha, \alpha)$$

Esse processo também é aplicado aos rótulos das classes, com o λ já escolhido, de modo que:

$$y_c = \lambda \cdot y_a + (1 - \lambda) \cdot y_b$$

Em nosso contexto, as instâncias X_i são as imagens e y_i o seu rótulo no formato *dummy* (*one-hot encoded*), ilustrado na figura 3.1.

(a) *Instância* X_a. $y_a = [1, 0]$ **(b)** *Instância* X_b. $y_b = [0, 1]$ **(c)** *Instância* X_c. $y_c = [0.4, 0.6]$

Figura 3.1: *MixUp aplicado em imagens de retina com* $\lambda = 0.4$. ***(a)*** *e* ***(b)*** *são originais do BRSET,* ***(c)*** *foi gerada com algoritmo descrito nesta seção.*
Fonte: *L. F. Nakayama et al., 2023*

Balanced MixUp

O Balanced MixUp (BMU) propõe estratégias de amostragens diferentes para as instâncias X_a, X_b e uma distribuição levemente diferente para λ. Neste algoritmo, as instâncias X_a e X_b são realizadas seguindo protocolos de amostragem baseados em instâncias (*instance-based sampling* - IBS) e em classes (*class-based sampling* - CBS), respectivamente.

Considere um conjunto $\mathcal{D}$ com N instâncias e K classes de cardinalidades n_j, isto é, $\sum_{j=1}^{K} n_j = N$. No IBS, temos que a probabilidade de uma instância X_a da classe j ser amostrada é igual a $P(X_a) = \frac{n_j}{N}$, já no CBS, as classes são equiprováveis e $P(X_a) = \frac{1}{K}$.

Explorações adicionais

Vamos avaliar a frequência esperada por classe (expected class frequency - ECF) seguindo o BMU.

Considere $y_{c,j}$ o valor da classe j da instância c.

$$ECF(n_j, \alpha) = \mathbb{E}[y_{c,j}] = \mathbb{E}[\lambda \cdot y_a + (1 - \lambda) \cdot y_b]$$
$$\Rightarrow ECF(n_j, \alpha) = \mathbb{E}[Beta(\alpha, 1)] \cdot \mathbb{E}[y_a] + \mathbb{E}[(1 - Beta(\alpha, 1)] \cdot \mathbb{E}[y_b]$$
$$\therefore ECF(n_j, \alpha) = \frac{\alpha}{\alpha + 1} \cdot \frac{n_j}{N} + \frac{1}{\alpha + 1} \cdot \frac{1}{K}$$

Note que o BMU tende a aproximar o ECF uniforme, afinal ele aplica uma interpolação convexa entre IBS e a CBS, no entanto um dataset com desbalanceamento muito agressivo ainda pode obter ECF desbalanceado.

Propomos aplicar entropia cruzada ponderada (WCE) com pesos proporcionais ao inverso da ECF após o BMU, ou seja:

$$w_j = \frac{1}{ECF_{BMU} \cdot K} = \frac{(\alpha + 1) \cdot N}{N + \alpha \cdot n_j \cdot K}$$

No subconjunto que usamos do BRSET, temos $n_j = [13320, 959]$, logo $ECF(0.1) = [0.53955266, 0.46044734]$ e $w = [1.07910533, 0.92089467]$.

3.4 Arquiteturas clássicas

Ao longo da evolução das CNNs, algumas arquiteturas específicas se destacaram, influenciando pesquisas subsequentes. Nesta seção, revisaremos brevemente algumas das arquiteturas clássicas, destacando suas contribuições.

3.4.1 LeNet-5

A LeNet-5, proposta por Yann LeCun et al. (LeCun *et al.*, 1998) é frequentemente citada como uma das primeiras CNNs bem-sucedidas em tarefas reais, especificamente o reconhecimento de dígitos escritos à mão no conjunto de dados MNIST. A rede era composta por camadas convolucionais, camadas de pooling (subamostragem) e camadas totalmente conectadas, culminando em uma camada de saída que previa a probabilidade de cada dígito. A LeNet-5 demonstrou a eficácia do uso de convoluções para extrair características relevantes de imagens, reduzindo a necessidade de engenharia manual de atributos.

Apesar de ser relativamente simples pelos padrões atuais, a LeNet-5 estabeleceu a base conceitual que muitas arquiteturas posteriores iriam seguir: uso de camadas convolucionais para extração hierárquica de características, pooling para redução espacial, funções de ativação não-lineares e treinamento end-to-end através do gradiente descendente.

3.4.2 AlexNet

A AlexNet, introduzida por Krizhevsky, Sutskever e Hinton em 2012, foi o ponto de inflexão que trouxe as CNNs para o centro das atenções em larga escala. Ao vencer o desafio ImageNet ILSVRC-2012 com uma margem significativa sobre as abordagens tradicionais, a AlexNet demonstrou que, com poder computacional suficiente e técnicas de regularização como dropout, era possível treinar redes convolucionais muito mais profundas e complexas.

A arquitetura da AlexNet apresentava oito camadas treináveis (cinco convolucionais e três totalmente conectadas), utilizava a função de ativação ReLU, aplicava pooling e empregava normalização de resposta local (LRN) (Krizhevsky *et al.*, 2012). Esse modelo inaugurou a era do aprendizado profundo em visão computacional, mostrando que a adição de mais camadas e o uso de grandes conjuntos de dados (como o ImageNet, com mais de um milhão de imagens) resultavam em melhorias drásticas de desempenho.

3.4.3 VGGNet

As VGGNets, introduzidas por Simonyan e Zisserman em 2014, exploraram a importância da profundidade da rede e do uso de filtros pequenos (3×3). A proposta principal foi simplificar a arquitetura, utilizando apenas convoluções com filtros muito pequenos, pooling e camadas totalmente conectadas ao final (Simonyan e Zisserman, 2014). Essa abordagem resultou em uma família de arquiteturas (VGG-11, VGG-13, VGG-16, VGG-19, número indicando o total de camadas) que foram bem-sucedidas no ImageNet.

As VGGNets demonstraram que, mantendo o tamanho do filtro fixo e pequeno e aumentando o número de camadas, era possível construir modelos altamente expressivos. A desvantagem das VGGNets é o elevado custo computacional e de memória devido ao grande número de parâmetros, especialmente nas camadas totalmente conectadas finais. Ainda assim, a clareza conceitual e o bom desempenho fizeram das VGGNets uma referência de arquitetura para muitos trabalhos subsequentes.

3.4.4 GoogLeNet (Inception)

A GoogLeNet, introduzida pela equipe do Google em 2014, apresentou o módulo Inception, cuja principal inovação foi aprender a observar a cena em múltiplas escalas ao mesmo tempo. Cada bloco Inception continha filtros de tamanhos diferentes (1×1, 3×3, 5×5), além de operações de pooling, todos concatenados no canal de saída. Dessa forma, a rede poderia simultaneamente extrair características locais finas e padrões mais globais, sem a necessidade de escolher manualmente o tamanho do filtro (Szegedy *et al.*, 2015).

A GoogLeNet era significativamente mais profunda (22 camadas) do que seus predecessores e, ainda assim, mais eficiente em termos computacionais, pois utilizava muitas convoluções 1×1 para reduzir a dimensionalidade, evitando explosão no número de parâmetros. Essa arquitetura demonstrou que a combinação de diferentes tamanhos de filtros, permitindo que a rede capturasse simultaneamente detalhes locais e estruturas globais através de diferentes campos receptivos, poderia melhorar a capacidade de representatividade da rede.

3.4.5 ResNet

As ResNets (Redes Residuais), introduzidas por He, Zhang, Ren e Sun em 2015, resolveram um problema crítico: conforme as redes se tornavam mais profundas, o treinamento tornava-se extremamente difícil devido à diminuição dos gradientes. As ResNets introduziram blocos residuais, que adicionam conexões diretas (skip connections) entre camadas não-adjacentes, permitindo que os gradientes fluíssem mais livremente através da rede (Kaiming He *et al.*, 2016). Isso possibilitou construir redes com dezenas ou mesmo centenas de camadas, melhorando ainda mais o desempenho.

A ideia central do bloco residual é que, ao invés de cada camada aprender uma transformação completa da entrada para a saída, ela aprende apenas a *diferença* (resíduo) em relação à entrada. Se uma camada não precisa mudar as características da entrada, ela pode simplesmente aprender uma função nula, o que é mais fácil do que aprender a função de identidade completa. Essa facilidade de otimização abriu caminho para arquiteturas

cada vez mais profundas e poderosas.

3.4.6 DenseNet

Posteriormente, muitas outras arquiteturas surgiram combinando ideias de convoluções eficientes, conexões densas e blocos de construção modulares. DenseNets (Huang *et al.*, 2017), por exemplo, ligam todas as camadas de forma densa, ajudando no fluxo de gradientes.

Dessa forma, a evolução das CNNs pode ser vista como uma contínua busca por representações mais ricas, redes mais profundas e eficientes, e soluções mais adequadas às restrições de hardware. Cada arquitetura clássica apresentou uma contribuição conceitual importante, incorporada posteriormente em novos modelos.

3.4.7 Vision Transformers (ViT)

Nos últimos anos, surgiu uma nova classe de arquiteturas que vem desafiando a hegemonia das CNNs em visão computacional: os **Vision Transformers** (ViT). Inspirados no sucesso dos Transformers em tarefas de processamento de linguagem natural, os ViTs foram introduzidos por Dosovitskiy *et al.*, 2021. A ideia fundamental é tratar uma imagem como uma sequência de **patches** (blocos de pixels) que são linearizados e depois processados por camadas Transformer, com mecanismos de **self-attention**.

Em vez de usar convoluções para extrair características, os ViTs baseiam-se em atenção global, permitindo que cada parte da imagem "se comunique" com todas as demais. Esse design pode capturar relações de longo alcance com mais facilidade do que as convoluções locais. Para tanto, cada **patch** é projetado em um vetor (**embedding**) e enriquecido com informações de posição (**positional embedding**). O Transformer processa essa sequência, retornando uma codificação que pode ser empregada para classificação, detecção ou outras tarefas de visão computacional.

Os ViTs exigem grandes quantidades de dados para serem bem treinados e, muitas vezes, são pré-treinados em conjuntos massivos (como o ImageNet-21k ou outros ainda maiores) e depois ajustados (**fine-tuned**) para tarefas específicas. Em diversos benchmarks, eles já alcançaram ou superaram o estado da arte das CNNs, inaugurando uma tendência de pesquisa cada vez maior em arquiteturas baseadas em atenção.

3.4.8 Swin Transformer

Na esteira dos ViTs, surgiram variações que buscam lidar melhor com alta resolução de imagens e reduzir o custo computacional. O **Swin Transformer** (**Shifted Window Transformer**), proposto em Liu *et al.*, 2021, é um exemplo marcante dessa evolução. Ele introduz janelas deslocadas (**shifted windows**) que particionam a imagem de forma hierárquica e processam localmente cada região, reduzindo a complexidade do mecanismo de **self-attention** em imagens grandes.

O Swin Transformer preserva a ideia de atenção, mas a aplica em blocos menores, recombinando gradualmente as informações. À medida que a rede avança, o tamanho das

janelas e o número de canais aumentam, criando uma arquitetura hierárquica semelhante às CNNs tradicionais em termos de extração multi-escala. Dessa forma, o Swin Transformer consegue aliar o melhor dos dois mundos: a flexibilidade do **self-attention** e a eficiência computacional típica das arquiteturas baseadas em escalas progressivas.

Capítulo 4

Trabalhos relacionados

As redes neurais convolucionais e os *Vision Transformers* (ViT), em particular, se mostraram ferramentas poderosas para a extração de características complexas das imagens de fundo de olho, permitindo identificar padrões associados à RD com alta precisão.

4.1 Categorias de estudos

Os estudos sobre RD podem ser divididos em duas categorias principais: detecção e graduação, conforme a tarefa. A detecção busca classificar imagens em dois grupos (retinas normais e com RD), enquanto a graduação procura identificar os estágios da doença.

Um levantamento realizado em 2023 identificou 53 artigos publicados entre 2017 e 2022 voltados para a tarefa de graduação da RD (Sebastian *et al.*, 2023).

Além disso, existe um corpo de estudo dedicado à segmentação das imagens de fundo de olho, classificando a imagem pixel a pixel. O grande diferencial deste tipo de abordagem é identificar a exata localização das lesões na imagem, o que contribui fortemente para a explicabilidade do eventual diagnóstico da doença.

4.2 Principais arquiteturas

Pesquisas recentes empregam arquiteturas como VGG (Jabbar *et al.*, 2022), (Abdel-Maksoud *et al.*, 2022), (Yaqoob *et al.*, 2021), ResNet (Martinez-Murcia *et al.*, 2021), (Butt *et al.*, 2022), Inception (AbdelMaksoud *et al.*, 2022), (Bilal *et al.*, 2022), EyeNet, DenseNet (AbdelMaksoud *et al.*, 2022) e EfficientNet (Canayaz, 2022), em geral utilizando os parâmetros pré-treinados para alavancar o aprendizado por transferência.

Além disso, métodos híbridos, que combinam CNNs com técnicas como SVMs ou módulos de atenção, também se destacam. Vários trabalhos exploram modelos híbridos dentro da própria camada convolucional, misturando arquiteturas diferentes na etapa de extração de features. Butt et. al, por exemplo, utiliza a estratégia de processar a imagem em duas redes diferentes: uma ResNet-18 e uma GoogleNet. Os vetores de features são mesclados e passados para o classificador (Butt *et al.*, 2022).

Modelos baseados em transformers dependem de um grande volume de dados anotados, o que costuma ser raro na área médica. No entanto, pesquisadores têm desenvolvido diversas estratégias para obter resultados competitivos em comparação com as CNNs em diversas aplicações da visão computacional no campo da medicina (Kelei He *et al.*, 2023).

Estudos recentes têm apresentado resultados promissores do uso de modelos baseados em transformers para detecção, graduação e segmentação da RD (Nazih *et al.*, 2023), (Mutawa e Sruthi, 2022). Nazih et al argumentam que o ViT pode ser superior às redes convolucionais, principalmente devido à sua capacidade de captar dependências globais. Essa vantagem decorre da estrutura de *multi-head self-attention*, em que cada pedaço da imagem "presta atenção"em todos os outros pedaços da imagem, que permite ao modelo identificar relações entre regiões distantes da imagem, superando as limitações estruturais das CNNs nesse aspecto.

Imagens de fundo de olho frequentemente apresentam lesões dispersas por toda a retina, como microaneurismas e áreas de hemorragia que podem estar distantes entre si. A *multi-head self-attention* do ViT possibilita que o modelo correlacione essas regiões distantes, captando padrões globais. Cada *head* de atenção calcula relações específicas entre diferentes partes da imagem. Essas informações são então combinadas para produzir uma compreensão mais holística da imagem.

Outra estratégia mais recente que têm mostrado potencial é a construção de modelos híbridos combinando CNNs e Transformers, aproveitando a capacidade das CNNs de extrair características locais e dos Transformers de capturar dependências globais. Por exemplo, (Sadeghzadeh *et al.*, 2023) utilizou o EfficientNet-B0 como extrator de características combinado com um Transformer para modelar dependências globais, alcançando resultados competitivos mesmo com conjuntos de dados menores. Outro trabalho (Saini *et al.*, 2023) propôs um modelo baseado em CNNs e Transformers para detectar e classificar estágios iniciais da RD, demonstrando uma compreensão holística das imagens de fundo de olho.

4.3 Conjuntos de dados

O sucesso dos modelos de DL depende fortemente da qualidade das bases de dados utilizadas. As mais utilizadas incluem Kaggle EyePACS (Dugas *et al.*, 2015), APTOS (Karthik *et al.*, 2019), MESSIDOR (Decencière *et al.*, 2014) e IDRiD (Porwal *et al.*, 2020). Essas bases variam em tamanho e qualidade, sendo amplamente utilizadas para validação e comparação de modelos. As principais características dos pricipais conjuntos de dados estão elencadas na tabela 4.1.

Conjunto de Dados	Tamanho	Anotações
Kaggle EyePACS	88.000+ imagens	Rotulado para detecção de retinopatia diabética (DR) e classificação por níveis de gravidade
Kaggle APTOS	5.590 imagens	Classificado por gravidade da DR (sem DR, leve, moderada, severa, proliferativa)
MESSIDOR	1.200 imagens	Presença e gravidade da DR (classificada)
IDRiD	516 imagens em alta resolução	Segmentação em nível de pixel para lesões (microaneurismas, hemorragias, exsudatos)
STARE	400 imagens	Anotado para segmentação de vasos sanguíneos e classificação da DR
DIARETDB1	89 imagens	Anotações em nível de lesão para microaneurismas, hemorragias e exsudatos

Tabela 4.1: *Resumo dos Conjuntos de Dados Públicos para Retinopatia Diabética*

Apesar dos avanços, desafios como o balanceamento de classes, a variabilidade das imagens de fundo de olho e a necessidade de interpretabilidade dos modelos permanecem.

há um interesse crescente em abordagens de IA explicável, que auxiliam profissionais médicos a confiarem mais nos sistemas de DL.

4.4 Preprocessamento

A maioria dos estudos citados no levantamento mencionado acima enfatizam a importância do preprocessamento na preparação de imagens de fundo de olho para tarefas de detecção e classificação da retinopatia diabética. Essas estratégias visam melhorar a qualidade das imagens, reduzir variabilidades causadas por diferentes dispositivos de captura e destacar características relevantes para os modelos de aprendizado profundo.

Uma das estratégias mais comuns é a melhoria da qualidade da imagem, que inclui uniformizar variabilidades relacionadas à iluminação, contraste e ruído introduzido durante a captura das imagens. Métodos como a equalização de histograma adaptativa limitada por contraste (CLAHE) são amplamente utilizados para melhorar o contraste local e aumentar a visibilidade de estruturas importantes, como os vasos sanguíneos e lesões (AbdelMaksoud *et al.*, 2022). Além disso, técnicas de remoção de ruído, como filtros Gaussianos e filtros medianos, ajudam a reduzir artefatos nas imagens, enquanto a normalização da intensidade ajusta os valores de pixel para minimizar variações de iluminação entre diferentes imagens (AbdelMaksoud *et al.*, 2022), (Bilal *et al.*, 2022), (Kaushik *et al.*, 2021), (Canayaz, 2022) e (Jabbar *et al.*, 2022).

Outra técnica importante é o ajuste de espaços de cor, que realça características específicas da retina ao priorizar canais de cor relevantes. Por exemplo, a extração do canal verde é amplamente adotada, pois fornece maior contraste para vasos sanguíneos e lesões, em comparação com os canais vermelho e azul (Boral e Thorat, 2021), (Suma e Kumar, 2018). A conversão para escala de cinza também é comum, pois reduz a complexidade computacional e a variabilidade de cores, ao mesmo tempo em que preserva informações estruturais essenciais (Boral e Thorat, 2021) e (AbdelMaksoud *et al.*, 2022).

O redimensionamento e o recorte das imagens são etapas essenciais para padronizar as dimensões das entradas nos modelos de aprendizado profundo. Imagens são frequentemente redimensionadas para resoluções fixas (como 224×224 ou 512×512) para se ajustarem às camadas de entrada de arquiteturas de CNNs mais utilizadas. Além disso, áreas irrelevantes ou de fundo são removidas por meio de recortes que focam especificamente na região da retina (AbdelMaksoud *et al.*, 2022), (Bilal *et al.*, 2022), (Canayaz, 2022), (Jabbar *et al.*, 2022), (Suma e Kumar, 2018).

Para enfrentar o problema do desbalanceamento de classes nos conjuntos de dados, muitos experimentos realizam o aumento de dados para produzir imagens alteradas por meio de técnicas como rotação, inversões, zoom e cisalhamento.

Capítulo 5

Metodologia

5.1 Definição da tarefa de classificação

As tarefas de classificação abordadas neste trabalho são:

- Detecção binária de RD
- Classificação do grau de severidade da RD

A detecção binária consiste em identificar se uma imagem de fundo de olho apresenta algum grau de retinopatia (1) ou se está saudável (0). Já a classificação multiclasses busca categorizar a imagem de acordo com uma escala de severidade da retinopatia, fornecendo informações mais detalhadas sobre a progressão da doença.

5.2 Conjunto de dados

Existem alguns conjuntos de dados públicos de imagens de fundo de olho.

5.2.1 EyePACS

O EyePACS Dataset é um repositório público amplamente utilizado em pesquisas e competições de detecção de RD, contendo 88.702 imagens de retina. As imagens foram capturadas em diferentes locais de atendimento médico, utilizando uma variedade de câmeras, o que reflete condições reais de captura. O conjunto inclui imagens dos olhos esquerdo e direito de pacientes, classificadas por profissionais treinados para classificar a RD em cinco níveis:

0. Sem RD
1. RD Leve
2. RD Moderada
3. RD Severa

4. RD Proliferativa

O dataset apresenta uma série de desafios, como a presença de ruído, que aparece nas imagens na forma de elementos visuais causados por reflexo, desfoque, sub ou superexposição, além da grande variabilidade nas condições de captura devido ao uso de diferentes câmeras e aparelhos. As imagens também apresentam desalinhamentos, com a mácula e o nervo óptico em posições diferentes.

Uma característica fundamental deste dataset é o forte desequilíbrio entre as classes (ver figura 5.1), já que a maioria das imagens é de olhos saudáveis, isto é, imagens classificadas como sem RD (classe 0). Esse fator faz com que estratégias de balanceamento sejam bastante importantes.

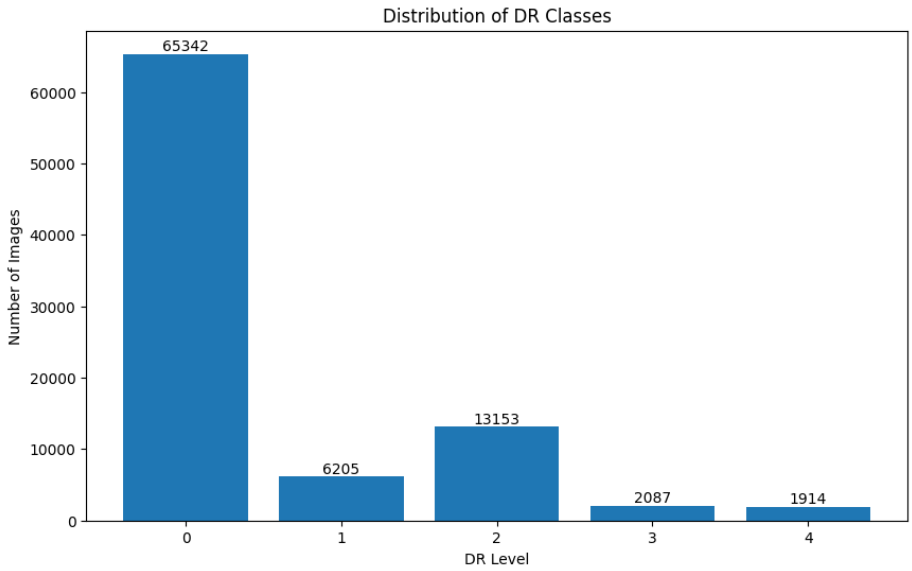


Figura 5.1: *Distribuição das classes dentro do EyePACS*
Fonte: *Produção própria.*

5.2.2 BRSET

O BRSET (Luis Filipe Nakayama *et al.*, 2024) é um conjunto público de dados que possui cerca de 16 mil fotos de retina. Essas imagens foram obtidas em três centros oftalmológicos em São Paulo entre 2010 e 2020. O objetivo do BRSET é abordar a escassez de conjuntos de dados públicos na região, aumentando a representatividade de dados e possibilitando a investigação de vieses algorítmicos em grupos demográficos diversos.

Além de conter as fotos de retina, o conjunto também inclui dados pessoais dos pacientes, como sexo, idade e histórico médico, anonimizados de forma que os pacientes não possam ser identificados.

Os instrumentos de captura foram câmeras Canon CR-2 e Nikon NF-505, principalmente.

As imagens são rotuladas por um especialista em retina, considerando características anatômicas (disco óptico, mácula), controle de qualidade (foco, iluminação) e patologias

(retinopatia diabética, edema macular, etc.).

Para determinação do grau de retinopatia diabética, o BRSET usa o padrão ICDR (Retinopatia Diabética Clínica Internacional) e o SDRG (Graduação de Retinopatia Diabética Escocês). De forma geral, esses padrões determinam os mesmos cinco graus para a RD que o EyePACS.

Para a tarefa de classificação por aprendizado de máquina, o BRSET apresenta dificuldades que precisam ser tratadas adequadamente, entre elas destacam-se imagens de baixa qualidade e o alto desbalanceamento entre fotos de retinas saudáveis e com RD. Cerca de 12% das imagens possuem alguma característica que diminui sua definição e só 6% das imagens apresentam algum grau de RD.

5.2.3 Synfundus

O SynFundus-1M (Shang *et al.*, 2024) é um dataset sintético composto por 1.000.018 imagens de fundo de olho, criado por meio do modelo de difusão probabilístico denominado SynFundus-Generator, o qual foi treinado com aproximadamente 1,3 milhão de imagens autênticas. Este modelo utiliza um autoencoder variacional para codificar imagens em um espaço latente comprimido e, posteriormente, aplica um modelo de difusão para gerar imagens sintéticas de alta qualidade.

Esse processo de geração resulta em amostras capazes de reproduzir, com alto grau de fidedignidade, as características visuais presentes em exames reais. Especialistas treinados com mais de cinco anos de experiência em algoritmos de imagens fundoscópicas têm dificuldade em distingui-las das imagens originais, com taxas de precisão próximas ao acaso. Além disso, o SynFundus-1M apresenta anotações confiáveis para 15 tipos de etiquetas, incluindo 11 doenças oculares (como retinopatia diabética, glaucoma, degeneração macular relacionada à idade e edema macular diabético) e 4 níveis de legibilidade de regiões fundamentais, como disco óptico, mácula e retina.

A legibilidade no contexto do SynFundus-1M refere-se à qualidade visual das imagens de fundo de olho e à capacidade de identificar, com clareza, regiões críticas para o diagnóstico, como o disco óptico, a mácula e a região retiniana. Essa qualidade é avaliada em quatro níveis: (1) a legibilidade global da imagem, indicando se ela é útil para qualquer julgamento clínico; (2) a legibilidade do disco óptico, que é essencial para avaliar condições como glaucoma; (3) a legibilidade da região retiniana, que abrange as áreas fora do disco óptico; e (4) a legibilidade da mácula, uma área fundamental para a visão central e detecção de doenças como degeneração macular. Fatores como artefatos, desfoco, subexposição ou superexposição podem comprometer essas regiões, afetando a precisão diagnóstica.

No SynFundus-1M, aproximadamente 91,5% das imagens são consideradas legíveis em todas as regiões críticas, garantindo que a maioria dos dados forneça informações confiáveis para análise clínica. Essa alta proporção de legibilidade não apenas assegura a consistência do treinamento de modelos, mas também reflete um esforço em simular as condições reais de exames oftalmológicos em um ambiente controlado e robusto.

A estratégia de "defesa"contra variabilidade em cenários clínicos considera a legibilidade em quatro níveis: se a imagem é globalmente legível (ou seja, útil para diagnóstico), e se as

regiões específicas (disco óptico, mácula e região retiniana) apresentam boa visibilidade. Imagens não legíveis são definidas por fatores como má focalização, subexposição ou superexposição que comprometem áreas críticas. No dataset, cerca de 91,5% das imagens são consideradas legíveis em todas as regiões, permitindo uma análise consistente de características fundamentais.

É importante notar que a forma de construção do SynFundus-1M exacerba as características mais frequentes de cada classe de doença, tornando o dataset mais "fácil" de ser aprendido pelos modelos. Isso significa que os sintomas mais típicos e visualmente evidentes, como hemorragias extensas ou margens indistintas do disco óptico, aparecem com maior frequência nas imagens sintéticas, enquanto características mais raras tendem a ser sub-representadas. Apesar disso, a riqueza de detalhes e a ampla cobertura do dataset o tornam extremamente útil para treinar modelos robustos.

O SynFundus-1M utiliza as mesmas cinco classes que o EyePACS. Entre as 1.000.018 imagens, há 804.442 rotuladas como negativas para retinopatia diabética (grau 0), enquanto 97.882 são de grau moderado (grau 2), 51.971 são severas (grau 3), e 45.667 apresentam retinopatia proliferativa (grau 4). Essa distribuição reflete a busca por um equilíbrio entre amostras positivas e negativas para melhorar o desempenho de modelos diagnósticos em doenças específicas.

A similaridade com as imagens reais é comprovada pela métrica Fréchet Inception Distance (FID), cujos valores indicam uma proximidade considerável entre os espaços latentes das imagens reais e sintéticas. A métrica é calculada comparando distribuições de características visuais extraídas por uma rede neural, e valores baixos significam que as diferenças entre os dois conjuntos de dados são mínimas. Isso assegura que os modelos treinados com imagens sintéticas generalizam bem para dados reais.

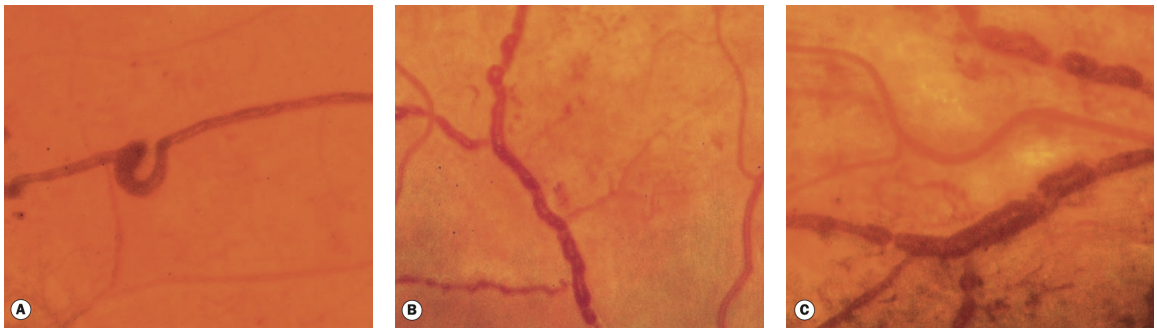
O SynFundus-1M tem se mostrado especialmente útil para o pré-treinamento de modelos na área de análise fundoscópica. Modelos pré-treinados com este dataset apresentam melhor desempenho e convergência mais rápida em tarefas de diagnóstico de doenças oculares, como a classificação de retinopatia diabética e glaucoma, em comparação a modelos pré-treinados em datasets genéricos como o ImageNet. Essa capacidade de aceleração do aprendizado e melhoria de performance reforça sua utilidade em pipelines de aprendizado profundo voltados à saúde.

O SynFundus-1M pode representar uma alternativa para superar a escassez de dados médicos em grande escala, mitigando riscos éticos associados à privacidade de pacientes. Ele é especialmente relevante para pesquisas em diagnóstico e classificação de doenças oculares, demonstrando que imagens sintéticas podem ser tão eficazes quanto dados reais para treinar modelos de ponta.

5.3 Preprocessamento das imagens

A retinopatia diabética é caracterizada pelas lesões descritas e ilustradas no capítulo 2. A RD é uma doença vascular e a observação da rede vascular da retina, isto é, do conjunto de vasos sanguíneos que irrigam a retina, é muito importante para o diagnóstico da doença e reconhecimento destas lesões.

Os vasos possuem forma tubular, com larguras e orientações variadas em uma mesma imagem, semelhantes às ramificações de uma raiz. A RD pode criar alterações venosas, caso em que pode ocorrer a bifurcação, *beading* e *looping*, ou segmentação dos vasos, com alteração da grossura e orientação dos vasos (Bowling, 2020, p. 528) (ver figura 5.2). Nesta hipótese, existe uma avaliação não só da aparência dos vasos em si, mas também da tortuosidade da rede vascular.



(a) *Looping* **(b)** *Beading* **(c)** *Segmentação*

Figura 5.2: *Alterações venosas*
Fonte: *(Bowling, 2020, P. 528)*

As lesões podem implicar também em alterações extra-venosas, como no caso dos microaneurismas, hemorragias e exsudatos, que aparecem na forma de pontos e manchas de colorações diferentes na imagem de retina.

Aplicamos algumas estratégias de preprocessamento que destacam os elementos visuais que caracterizam as lesões. As etapas realizadas estão descritas abaixo:

1. **Centralização da Retina:** Utilizamos a **Transformada de Hough** para identificar o círculo que representa a retina, conforme Figura 5.3. A imagem foi então centralizada, garantindo que a retina estivesse posicionada no centro do quadro. Essa etapa é essencial para lidar com imagens desalinhadas ou mal capturadas, melhorando a consistência entre as imagens de entrada.

2. **Realce de Detalhes:** Para destacar elementos importantes, aplicamos um filtro Gaussiano para criar uma versão borrada (*blurry*) da imagem e subtraímos os valores de pixel dessa versão da imagem original. Essa técnica, que **remove a média local de cores**, ressalta estruturas de interesse, eliminando variações causadas por iluminação desigual.

3. **Aplicação de Máscara:** Criamos uma **máscara circular** para excluir a borda da retina, que frequentemente apresenta grande contraste com o restante da imagem. Essa borda poderia introduzir ruído e interferir na análise, por isso, sua exclusão aumenta a probabilidade de o modelo se concentrar nos elementos visuais mais relevantes: as lesões.

4. **Correção do Raio da Retina:** Ajustamos o raio da retina para padronizar a escala entre as imagens. Essa etapa facilita a análise visual e garante que todas as imagens apresentem dimensões consistentes para o modelo.

5. **Normalização e Redimensionamento:** Redimensionamos as imagens para 512 × 512 pixels e normalizamos seus valores conforme a média e desvio padrão calculados no dataset inteiro. Essas etapas garantem uniformidade no conjunto de dados e aceleram o treinamento.

Todas as etapas foram implementadas utilizando bibliotecas como `PyTorch` e `OpenCV`, com suporte a processamento paralelo para otimização. As técnicas descritas foram baseadas em estudos prévios e benchmarks realizados com o dataset EyePACS. Pode-se visualizar exemplos de resultados nas figuras 5.4, 5.5 e 5.6.

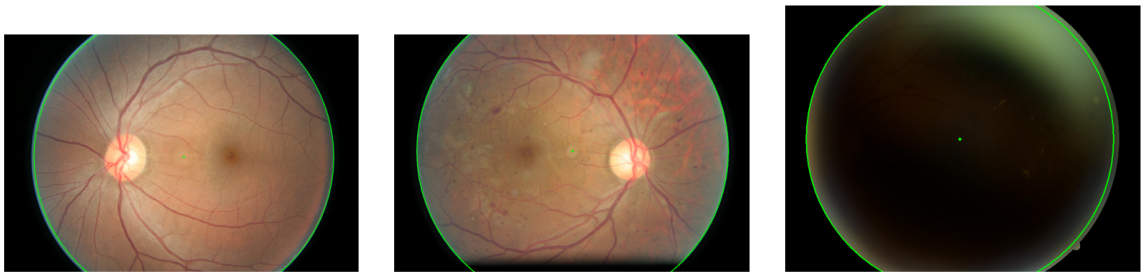


Figura 5.3: *Aplicação da Transformada de Hough em imagens de diferentes condições, indicados pela circunferência e ponto central verdes.*
Fonte: *Dugas et al., 2015*

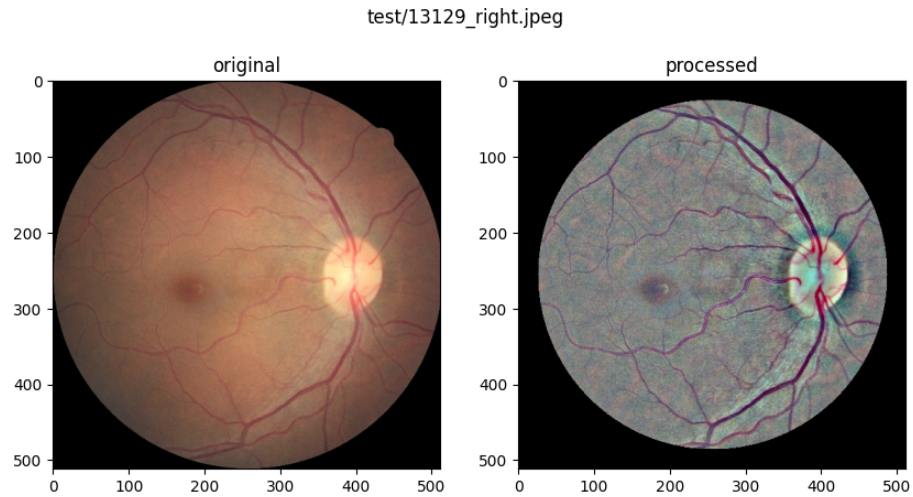


Figura 5.4: *Exemplo 1 de imagem preprocessada seguindo algoritmo descrito em 5.3*
Fonte: *Produção própria.*

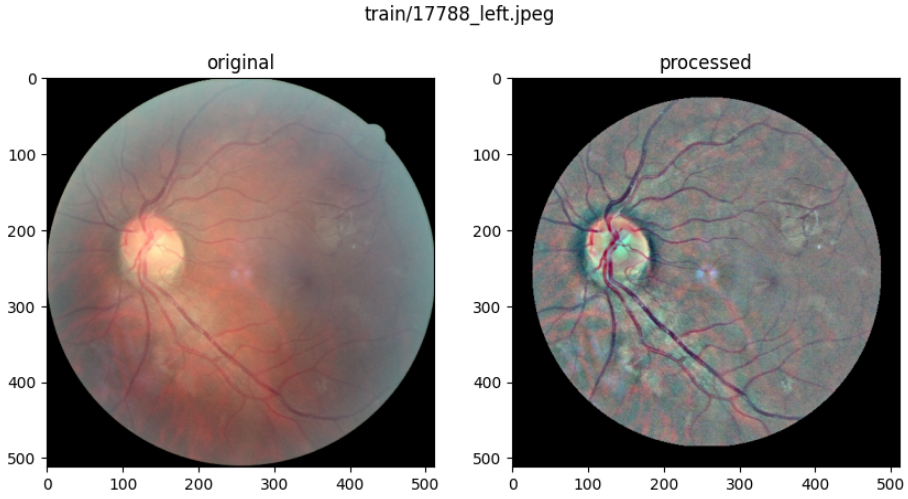


Figura 5.5: *Exemplo 2 de imagem preprocessada seguindo algoritmo descrito em 5.3*
Fonte: *Produção própria.*

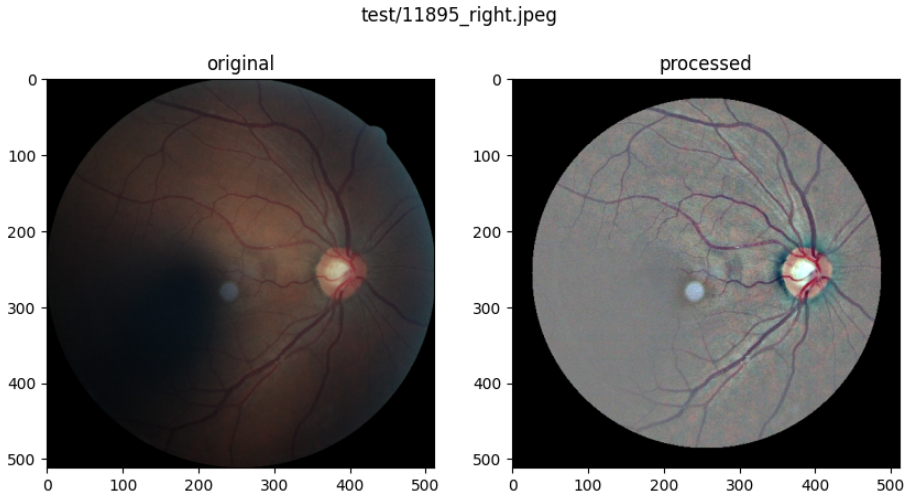


Figura 5.6: *Exemplo 3 de imagem preprocessada seguindo algoritmo descrito em 5.3*
Fonte: *Produção própria.*

5.4 Aumento de dados

Para aumentar a robustez do modelo e aprimorar sua capacidade de generalização, aplicamos técnicas de aumento de dados ao conjunto de treinamento, utilizando uma combinação de transformações que simulam variações naturais nas imagens.

Inicialmente, realizamos ajustes que incluem rotações aleatórias de até 30 graus, deslocamentos horizontais e verticais de até 10% do tamanho da imagem, variações de escala entre 90% e 110%, além de distorções que alteram o ângulo das imagens em até 18 graus. Em seguida, aplicamos uma inversão horizontal aleatória, realizada com uma probabilidade de 50%. Essas técnicas foram selecionadas para reproduzir variações comuns nas condições de

captura da imagem, como mudanças de perspectiva e orientação, aumentando a diversidade do conjunto de dados e reduzindo o risco de overfitting.

5.5 Hiperparâmetros

O modelo foi inicializado com pesos pré-treinados (`pretrained` configurado como `true`) e camadas de *batch normalization* foram incorporadas para melhorar a estabilidade e a convergência durante o treinamento. A camada totalmente conectada oculta foi configurada com 1024 neurônios, e uma taxa de dropout de 0,2 (`p_dropout` = 0,2) foi aplicada para mitigar o overfitting.

O processo de carregamento dos dados foi configurado com um `batch_size` de 32, e o embaralhamento dos dados foi ativado (`shuffle` = `true`) para garantir uma distribuição aleatória entre as amostras de cada lote. Além disso, utilizamos 4 *workers* para otimizar a leitura e o processamento dos dados, reduzindo gargalos durante o treinamento.

O otimizador escolhido foi o **Adam**, configurado com uma taxa de aprendizado inicial de 1×10^{-4}, conhecido por sua eficiência em problemas com grande dimensionalidade. A função de perda adotada foi a entropia cruzada (`cross_entropy_loss`), amplamente utilizada para tarefas de classificação, enquanto a métrica de desempenho monitorada durante o treinamento foi a acurácia.

Para o ajuste dinâmico da taxa de aprendizado, implementamos um `scheduler` do tipo `ReduceLROnPlateau`, configurado com um fator de redução de 0,9, paciência de 5 épocas e uma taxa de aprendizado mínima (`min_lr`) de 1×10^{-6}. Este ajuste contribuiu para uma convergência mais eficiente, reduzindo o risco de estagnação do modelo.

Adicionalmente, utilizamos o TensorBoard para a visualização em tempo real das métricas de desempenho, como acurácia e perda, permitindo o acompanhamento detalhado do progresso do treinamento e auxiliando na identificação de problemas como overfitting ou aprendizado lento.

5.6 Avaliação

Nesta seção, detalhamos as métricas utilizadas para avaliar o desempenho do modelo em relação à tarefa de classificação. A escolha dessas métricas é justificada pela necessidade de considerar diferentes aspectos da qualidade das previsões, como proporção de acertos, equilíbrio entre classes e concordância ponderada.

5.6.1 Precisão

A precisão mede a proporção de exemplos classificados como positivos que realmente pertencem à classe positiva. É calculada como:

$$\text{Precisão} = \frac{\text{VP}}{\text{VP} + \text{FP}}$$

Onde: - **VP** (Verdadeiros Positivos): Exemplos corretamente classificados como pertencentes à classe positiva. - **FP** (Falsos Positivos): Exemplos incorretamente classificados como pertencentes à classe positiva.

Esta métrica é particularmente útil quando os erros de falsos positivos possuem maior impacto no problema abordado.

5.6.2 Acurácia

A acurácia é uma métrica global que mede a proporção de exemplos corretamente classificados entre todos os exemplos do conjunto de dados. Sua fórmula é dada por:

$$\text{Acurácia} = \frac{\text{Total de acertos}}{\text{Total de exemplos}}$$

Embora seja amplamente utilizada, a acurácia pode ser uma métrica insuficiente para problemas com classes desbalanceadas, pois não leva em conta a distribuição dos dados entre as classes.

5.6.3 Revocação

A revocação, também chamada de sensibilidade ou taxa de verdadeiro positivo, mede a proporção de exemplos da classe positiva que foram corretamente identificados pelo modelo. É calculada como:

$$\text{Revocação} = \frac{\text{VP}}{\text{VP} + \text{FN}}$$

Onde:

- VP (Verdadeiros Positivos): Exemplos corretamente classificados como pertencentes à classe positiva.
- FN (Falsos Negativos): Exemplos da classe positiva que foram incorretamente classificados como pertencentes à classe negativa.

Essa métrica é especialmente relevante em problemas onde minimizar os falsos negativos é mais importante, como no diagnóstico médico, onde não identificar uma doença pode ter consequências graves.

5.6.4 F1-Score

O F1-score é a média harmônica entre a precisão e a revocação, oferecendo uma métrica equilibrada que leva em conta tanto falsos positivos quanto falsos negativos. Ele é calculado como:

$$\text{F1-Score} = 2 \cdot \frac{\text{Precisão} \cdot \text{Revocação}}{\text{Precisão} + \text{Revocação}}$$

O F1-score é útil em cenários onde há um desequilíbrio entre as classes, pois fornece uma visão mais equilibrada do desempenho do modelo em relação à precisão e revocação. Valores mais próximos de 1 indicam que o modelo está alcançando um bom equilíbrio entre as duas métricas.

5.6.5 Quadratic Weighted Kappa

O quadratic weighted kappa (ou kappa quadrado ponderado) é uma métrica de concordância que avalia a consistência entre as previsões do modelo e os rótulos reais, ponderando os erros com base em sua gravidade. Sua fórmula é baseada no coeficiente de kappa de Cohen, mas adiciona pesos para considerar a magnitude do erro, o que é especialmente relevante para tarefas de classificação em que existe uma graduação.

No contexto da classificação do nível da RD, este tipo de medida permite ponderar os resultados pela distância entre a previsão e a classe esperada. Por exemplo, se a classe esperada para uma determinada amostra era quatro, mas o modelo classificou como três, este erro deve ser menos importante para a aferição da performance do modelo que se a classe prevista fosse zero, não obstante ambas as previsões estejam incorretas.

Matematicamente, é definido como:

$$\kappa = 1 - \frac{\sum_{i=1}^{N} \sum_{j=1}^{N} W_{ij} O_{ij}}{\sum_{i=1}^{N} \sum_{j=1}^{N} W_{ij} E_{ij}}$$

Onde:

- N: número total de categorias.
- W_{ij}: matriz de pesos, que penaliza os erros com base na distância quadrática entre as categorias i e j. É calculada como:

$$W_{ij} = \frac{(i-j)^2}{(N-1)^2}$$

- O_{ij}: matriz de confusão observada, representando o número de vezes que a classe real i foi prevista como j.
- E_{ij}: matriz de confusão esperada, que seria a matriz de frequências esperadas assumindo que as classificações foram feitas ao acaso. É calculada como:

$$E_{ij} = \frac{\text{sum\_row}_i \cdot \text{sum\_col}_j}{\text{total}}$$

  Onde sum_row_i é o total de vezes que a classe i aparece na verdade, e sum_col_j é o total de vezes que a classe j foi prevista.

Interpretação:

- $\kappa = 1$: Concordância perfeita.

- $\kappa = 0$: Concordância equivalente à esperada ao acaso.
- $\kappa < 0$: Discordância pior do que o acaso.

O QWK é especialmente útil em tarefas onde as previsões erradas têm gravidades diferentes, valorizando previsões que estejam mais próximas da classe correta.

5.7 Desbalanceamento

Para lidar com o desbalanceamento do BRSET, experimentamos algumas abordagens: entropia cruzada ponderada, a técnica Balanced MixUp descrita no capítulo 3, a aplicação de ambos ao mesmo tempo com estratégias de pesos diferentes.

Capítulo 6

Experimentos e resultados

Dividimos os experimentos em duas classes, o problema de graduação de severidade e o de detecção, referentes à classificação de retinopatia diabética em graus e binário.

Os modelos foram treinados no Laboratório e-Science do Instituto de Matemática e Estatística da USP em máquinas com processador Intel(R) Xeon(R) Silver 4110 CPU @ 2.10GHz 32 cores e GPUs NVIDIA GeForce GTX TITAN X de 12GB.

Os códigos de treinamento foram baseados em um template de treinamento em pytorch de código aberto disponível https://github.com/victoresque/pytorch-template. A nossa versão, adequada para o problema de retinopatia diabética, está disponível em https://github.com/lucasmsobrinho/dr-pytorch-training-pipelines.

6.1 Detecção de RD

Para o problema de detecção, utilizamos o conjunto de dados BRSET com pré-processamento baseado na transformação de Hough para localização da circunferência, seguida do redimensionamento das imagens para 224x224 pixels. O conjunto foi dividido em 70% para treinamento, 10% para validação (utilizada apenas para early stopping) e 20% para teste. Optamos por utilizar os modelos Swin tiny, ViT small e ResNet50, por serem representantes de técnicas estado da arte e apresentarem quantidade de parâmetros comparáveis entre si.

Experimento 1: Baseline

O primeiro experimento consistiu na realização do fine-tuning dos modelos pré-treinados no ImageNet, seguindo uma configuração padronizada de hiperparâmetros. Foi utilizado o otimizador Adam, com taxas de aprendizado ajustadas para cada modelo: 10^{-5} para a ResNet50, 10^{-6} para o modelo Swin e 5×10^{-6} para o ViT. O treinamento foi conduzido com um batch size de 32 e por até 50 épocas, sendo o modelo final selecionado com base na menor perda de validação.

Experimentos 2 a 5: Tratamento de Desbalanceamento

Com o objetivo de mitigar o desbalanceamento presente na classificação do BRSET, quatro abordagens distintas foram avaliadas. A primeira abordagem utilizou Weighted Cross Entropy (WCE), aplicando uma estratégia comum de pesos (inverso da frequência da classe). Em seguida, testou-se o método Balanced Mix Up (BMU), detalhado em 3.3.1. Como terceira estratégia, foi realizada a combinação do BMU com WCE indicado por BMU+WCE. Por fim, o método proposto em 3.3.1, indicado como BMU+WCE+ECF, e que usamos como peso do WCE a frequência esperada das classes pós-BMU.

No segundo experimento - com a adição de WCE - as taxas de aprendizado usadas foram as mesmas do experimento 1, exceto para o ViT, onde foi usada uma taxa de 10^{-6}. No experimento 3, onde usou-se BMU pela primeira vez, as taxas de aprendizado foram 10^{-6} para os modelos Transformer e 5×10^{-5} para a ResNet50, assim como no experimento 4 - onde BMU e WCE foram aplicados simultaneamente. No experimento 5, com a combinação BMU+WCE+ECF, a taxa da ResNet50 foi de 10^{-5} e as outras foram mantidas iguais.

Os outros hiperparâmetros não mencionados foram mantidos iguais aos do Experimento 1.

Experimentos 6 e 7: Pré-treinamento com SynFundus-1M

Com a intenção de aprimorar o desempenho dos modelos, realizamos um pré-treinamento utilizando o conjunto de dados SynFundus-1M, com uma seleção de 250 mil imagens de qualidade maior que 7. Esse dataset foi dividido em conjuntos de Treino e Validação com 240 mil e 10 mil instâncias, respectivamente. Durante essa etapa, foram empregados os seguintes hiperparâmetros: batch size de 64, otimizador Adam com taxa de aprendizado constante de 3×10^{-6}. O treinamento foi conduzido por 5 épocas, com os dados balanceados em 125 mil imagens saudáveis e 125 mil com retinopatia diabética.

Na etapa de fine-tuning, foram realizados dois experimentos distintos. O primeiro consistiu no fine-tuning baseline, utilizando o Adam com taxas de aprendizado de 3×10^{-5} para Swin e ViT e $3 \cdot 10^{-4}$ para ResNet por 25 épocas. No segundo experimento aplicamos o método proposto (BMU+WCE+ECF), mantendo os mesmos hiperparâmetros do fine-tuning baseline.

6.2 Graduação de severidade

Para treinamento desse modelo, utilizamos o conjunto de dados EyePacs divido em 28107 imagens de treino e 6984 de validação (80:20), aplicando o algoritmo de localização de circunferências baseado na transformação de Hough para centralizar as retinas e redimensioná-las para 512x512. Por fim, o filtro de realce de detalhes a partir da subtração local de cor descrito na seção 5.3.

Usamos a VGG16 pré-treinada ImageNet em sua versão com Batch Normalization. Com otimizador SGD e learning scheduler Cyclic Learning Rate oscilante entre 10^{-2} até 10^{-4}.

6.3 Resultados e discussões

a) Detecção de RD

Modelo	Experimento	Acurácia (%)	Macro Precisão (%)	Macro Revocação (%)	F1-Score (%)
ResNet	1) Baseline	94,92	85,15	67,46	75,28
	2) WCE	90,26	68,42	84,52	75,62
	3) BMU	95,30	82,62	76,47	79,42
	4) BMU+WCE	91,94	71,12	84,93	77,41
	5) BMU+WCE+ECF	95,02	80,08	79,49	79,79
	6) Pré-treino + Baseline	96,24	90,80	76,24	82,88
	7) Pré-treino + BMU+WCE+ECF	96,49	90,35	79,06	84,33
ViT	1) Baseline	95,41	85,35	73,10	78,76
	2) WCE	86,54	64,57	83,99	73,01
	3) BMU	94,92	79,63	79,19	79,41
	4) BMU+WCE	87,59	65,38	83,82	73,46
	5) BMU+WCE+ECF	94,78	78,90	79,61	79,25
	6) Pré-treino + Baseline	95,12	84,69	70,50	76,95
	7) Pré-treino + BMU+WCE+ECF	95,69	89,06	72,27	79,79
Swin	1) Baseline	97,23	**91,64**	84,83	88,10
	2) WCE	90,68	70,02	90,12	78,81
	3) BMU	96,25	83,29	90,17	86,60
	4) BMU+WCE	92,04	72,31	**92,07**	81,00
	5) BMU+WCE+ECF	93,34	83,33	90,68	86,85
	6) Pré-treino + Baseline	**97,51**	91,19	88,15	**89,65**
	7) Pré-treino + BMU+WCE+ECF	97,26	90,73	86,31	88,47

Tabela 6.1: *Resultados dos treinamentos avaliados no conjunto de testes. Os melhores resultados por arquitetura estão sublinhados, destacando em negrito os melhores dentre todos.*

Observou-se em 6.1 que a estratégia de Entropia Cruzada Ponderada (WCE) com pesos inversos à frequência de classe resultou em um menor f1-score comparado aos baselines dos modelos baseados em Transformers e um resultado bem próximo no caso da ResNet.

Isso ocorre por que o desbalanceamento das classes nesse dataset é bastante intenso, havendo apenas 6% de classes positivas, logo o peso atribuído à classe sub-representada no WCE fica muito alto e enviesando o treinamento nesse sentido. Esse resultado é reiterado pela revocação consideravelmente maior do que a precisão no experimento 2 em todos os modelos, indicando um sobreajuste na classe positiva.

Esse resultado também é observado no experimento 4 (BMU+WCE) em comparação com o experimento 3 (BMU).

O experimento 3, que testa a implementação do Balanced MixUp, superou o baseline com os modelos ViT e ResNet, mas não com o Swin.

Já a implementação do BMU com ECF, relativo ao experimento 5, não apresentou melhoras significativas frente às métricas avaliadas nos treinamentos executados.

O pré-treino usando o SynFundus-1M demonstrou ser uma abordagem robusta para aumentar o desempenho, com benefícios claros para ResNet e Swin, reforçando a importância de aproveitamento de conhecimento prévio no treinamento de redes neurais.

Por fim, o Swin obteve melhores resultados gerais de f1-score e acurácia em comparação com os outros modelos, em especial com o finetuning após o pré-treino no SynFundus-1M, que obteve o maior f1-score (89,65%) e acurácia (97,51%) dentre todos os experimentos.

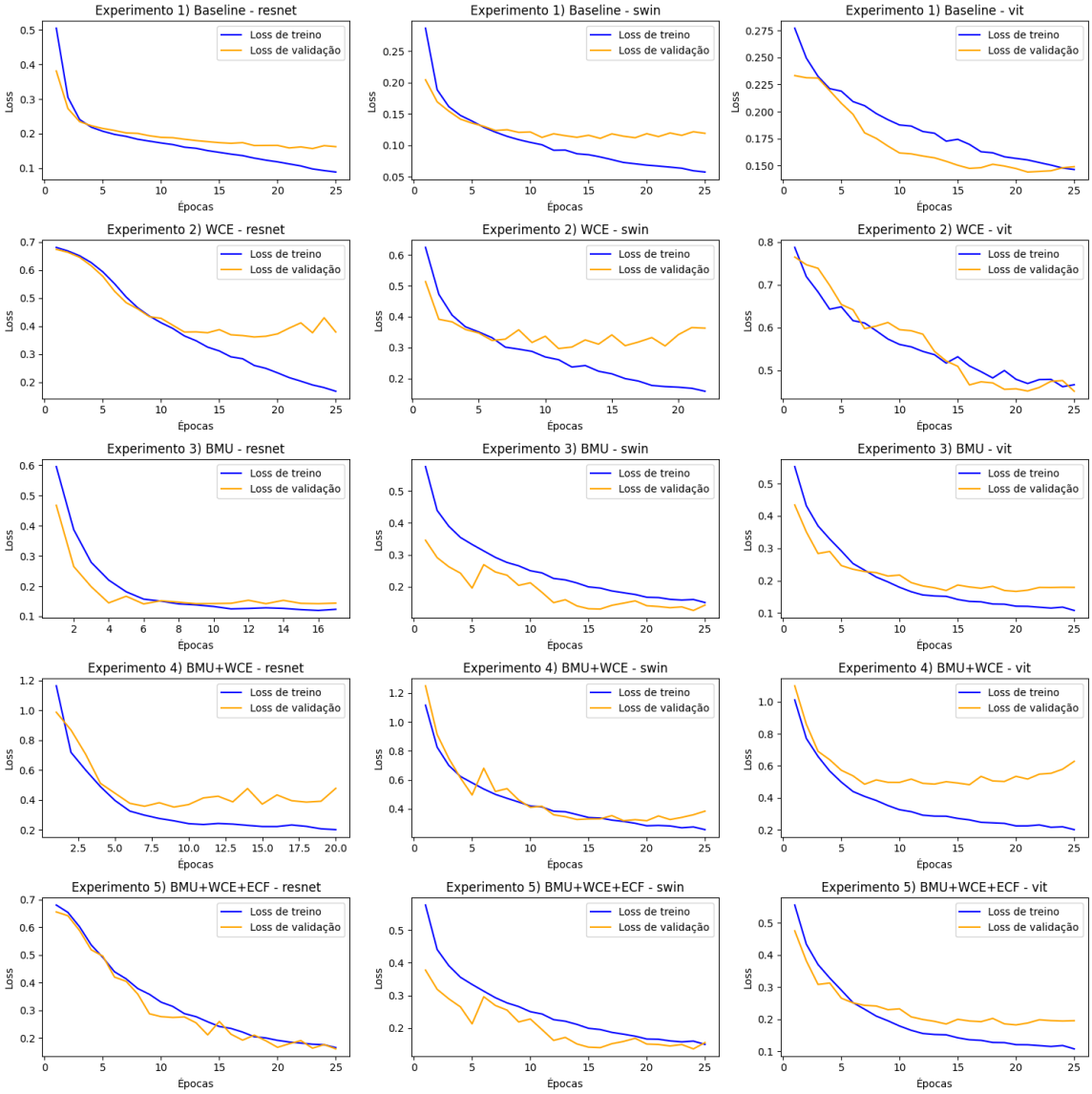


Figura 6.1: *Gráficos de loss de treino de validação dos experimentos 1) a 5).*
Fonte: *Produção própria.*

Pontos de Melhoria

Os resultados obtidos nos treinamentos indicaram limitações que devem ser consideradas para aprimorar futuros experimentos. Os gráficos de loss da figura 6.1 sugerem que os modelos atingem seu menor valor de perda rapidamente, sem melhorias significativas nas épocas subsequentes, além disso as curvas apresentam variações abruptas entre as épocas. Isso pode ser resultado de hiperparâmetros sub-ótimos, como a taxa de aprendizado e necessidade de métodos adicionais de regularização. Por fim, a quantidade e diversidade dos dados de treinamento podem não ter sido o suficiente para explorar o potencial completo das arquiteturas avaliadas.

Abordagens Possíveis

Para abordar as limitações observadas, futuros trabalhos podem implementar uma busca sistemática de hiperparâmetros mais robusta, como *grid search* ou otimização bayesiana, a fim de identificar configurações ideais para cada modelo. Implementar outros métodos de regularização mais agressivos pode ser útil. Também seria benéfico expandir o conjunto de dados de treinamento com outros conjuntos públicos disponíveis, fortificando a representatividade e a robustez dos modelos.

b) Graduação de severidade

Foi atingido um resultado de acurácia: 83,35%, com macro precisão de 64,02% e macro revocação de 51,06%, totalizando um f1-score de 56,81% e quadratic weighted kappa de 0,72.

Capítulo 7

Conclusão

Este trabalho resultou no desenvolvimento de um modelo para a detecção de retinopatia diabética a partir de imagens de fundo de olho, explorando diferentes arquiteturas de aprendizado de máquina.

A pesquisa envolveu a análise dos conjuntos de dados disponíveis, analisando suas características, limitações e desafios inerentes ao problema. O projeto incluiu também o estudo da caracterização visual da RD, permitindo compreender melhor as variações presentes nas imagens e sua influência no desempenho dos modelos.

Outro aspecto fundamental deste trabalho foi a investigação das métricas utilizadas para avaliar modelos de classificação, permitindo uma análise apurada dos resultados. Foram exploradas métricas como acurácia, F1-score, macro precisão e macro revocação, considerando a importância de cada uma no contexto de um problema com classes desbalanceadas. Além disso, aprofundou-se o estudo sobre arquiteturas modernas de aprendizado de máquina, compreendendo seus princípios de funcionamento e como diferentes abordagens podem impactar a tarefa de classificação.

Por fim, foi realizada uma avaliação criteriosa de técnicas de preprocessamento e de estratégias para lidar com o desbalanceamento dos dados, fatores essenciais para aprimorar a generalização do modelo. Foram experimentadas diferentes abordagens, analisando sua eficácia na melhoria do desempenho do classificador. Os resultados obtidos demonstraram a importância dessas etapas no pipeline de aprendizado de máquina, reforçando a necessidade de um tratamento cuidadoso dos dados antes do treinamento dos modelos.

Mais concretamente, os principais frutos desta pesquisa incluem o treinamento de um modelo de detecção automática de retinopatia diabética, que alcançou 97,5% de acurácia e 89,5% de F1-score no conjunto de testes do BRSET, com o uso da arquitetura Swin Transformer, mas também foram investigados o Vision Transformer (ViT) e a ResNet. Além disso, explorou-se técnicas para lidar com conjuntos de dados desbalanceados, em que se reimplementou e validou o BalancedMixUp no contexto da retinopatia diabética. Como contribuição adicional, foi proposto o uso do BMU em conjunto com a Entropia Cruzada Ponderada (WCE) aplicando uma estratégia de pesos própria, o que demonstrou potencial para mitigar ainda mais os problemas de treinamento em conjuntos desbalanceados.

Os desenvolvimentos apresentados neste estudo abrem espaço para diversas direções futuras de pesquisa. Uma possibilidade é investigar o impacto de diferentes ajustes no parâmetro α utilizado no cálculo do BMU, avaliando seu efeito sobre a performance geral dos modelos. Além disso, a exploração de abordagens semelhantes, como o método Remix (Chou *et al.*, 2020), pode fornecer novas perspectivas para lidar com desafios relacionados ao balanceamento de classes e generalização.

Uma tarefa adicional importante é a aplicação das técnicas desenvolvidas neste trabalho em outros conjuntos de dados, como o EyePACS, para avaliar a sua efetividade em diferentes contextos e problemas reais.

Também seria interessante experimentar métodos alternativos de preprocessamento das imagens, como a subtração da cor local média.

Em suma, as explorou-se o uso de técnicas de aprendizado de máquina para a detecção da retinopatia diabética, destacando a importância de técnicas adequadas de preprocessamento, balanceamento de dados e escolha de arquiteturas. Os resultados obtidos reforçam a possibilidade de usar a inteligência artificial como uma ferramenta no apoio à clínica médica.

Referências

[AbdelMaksoud *et al.* 2022] Eman AbdelMaksoud, Sherif Barakat e Mohammed Elmogy. "A computer-aided diagnosis system for detecting various diabetic retinopathy grades based on a hybrid deep learning technique". *Medical & Biological Engineering & Computing* 60 (2022), pp. 2015–2038. doi: 10.1007/s11517-022-02564-6 (citado nas pgs. 21, 23).

[American Academy of Ophthalmology 2021] American Academy of Ophthalmology. *International Clinical Classification System for Diabetic Retinopathy and Diabetic Macular Edema*. Accessed: 2024-11-09. 2021. url: https://www.aao.org/education/clinical-statement/international-clinical-classification-system-diabe (citado na pg. 3).

[ASRS s.d.] Image Banks ASRS. *Proliferative diabetic retinopathy with vitreous hemorrhage: Color fundus photos*. Accessed: 2025-01-18. url: https://imagebank.asrs.org/file/1553/proliferative-diabetic-retinopathy-with-vitreous-hemorrhage-color-fundus-photos (citado nas pgs. 6, 8).

[Bilal *et al.* 2022] A. Bilal, L. Zhu, A. Deng, H. Lu e N. Wu. "Ai-based automatic detection and classification of diabetic retinopathy using u-net and deep learning". *Symmetry* 14.1427 (2022). doi: 10.3390/symmetry14071427 (citado nas pgs. 21, 23).

[Boral e Thorat 2021] Prajakta Boral e Siddhi Thorat. "Classification of diabetic retinopathy based on hybrid neural network". *International Journal of Engineering Research and Technology* 10.10 (2021), pp. 237–243 (citado na pg. 23).

[Bowling 2020] Brad Bowling. *Kanski's Clinical Ophthalmology: A Systematic Approach*. 9th. Elsevier, 2020. isbn: 9780702077111 (citado nas pgs. v, 3, 5–7, 29).

[Butt *et al.* 2022] Muhammad Mohsin Butt, D. N. F. Awang Iskandar, Sherif E. Abdelhamid, Ghazanfar Latif e Runna Alghazo. "Diabetic retinopathy detection from fundus images of the eye using hybrid deep learning features". *Diagnostics* 12.7 (2022). Licensee MDPI, Basel, Switzerland. Distributed under CC BY license., p. 1607. doi: 10.3390/diagnostics12071607. url: https://www.mdpi.com/article/10.3390/diagnostics12071607 (citado na pg. 21).

[Canayaz 2022] M. Canayaz. "A hybrid feature extraction method with deep learning for diabetic retinopathy classification". *Applied Soft Computing* 125 (2022), p. 109462. doi: 10.1016/j.asoc.2022.109462 (citado nas pgs. 21, 23).

[Chagas *et al.* 2023] Thiago Alves Chagas *et al.* "Prevalence of diabetic retinopathy in brazil: a systematic review with meta-analysis". *Diabetology & Metabolic Syndrome* 15.1 (2023), p. 34. doi: 10.1186/s13098-023-01003-2. url: https://doi.org/10.1186/s13098-023-01003-2 (citado na pg. 9).

[Chou *et al.* 2020] Hsin-Ping Chou, Shih-Chieh Chang, Jia-Yu Pan, Wei Wei e Da-Cheng Juan. *Remix: Rebalanced Mixup*. 2020. arXiv: 2007.03943 [cs.CV]. url: https://arxiv.org/abs/2007.03943 (citado na pg. 44).

[Decencière *et al.* 2014] Etienne Decencière *et al.* "Feedback on a publicly distributed database: the messidor database". *Image Analysis & Stereology* 33.3 (ago. de 2014), pp. 231–234. issn: 1854-5165. doi: 10.5566/ias.1155. url: http://www.ias-iss.org/ojs/IAS/article/view/1155 (citado na pg. 22).

[Dosovitskiy *et al.* 2021] Alexey Dosovitskiy *et al.* "An image is worth 16x16 words: transformers for image recognition at scale". In: *International Conference on Learning Representations (ICLR)*. 2021. url: https://arxiv.org/abs/2010.11929 (citado na pg. 19).

[Duchi *et al.* 2011] John Duchi, Elad Hazan e Yoram Singer. "Adaptive subgradient methods for online learning and stochastic optimization". *J. Mach. Learn. Res.* 12.null (jul. de 2011), pp. 2121–2159. issn: 1532-4435 (citado na pg. 14).

[Dugas *et al.* 2015] Emma Dugas, Jared, Jorge e Will Cukierski. *Diabetic Retinopathy Detection*. https://kaggle.com/competitions/diabetic-retinopathy-detection. Kaggle. 2015 (citado nas pgs. 22, 30).

[Galdran *et al.* 2021] Adrian Galdran, Gustavo Carneiro e Miguel A. González Ballester. "Balanced-mixup for highly imbalanced medical image classification". In: *Medical Image Computing and Computer Assisted Intervention – MICCAI 2021*. Springer International Publishing, 2021, pp. 323–333. isbn: 9783030872403. doi: 10.1007/978-3-030-87240-3_31. url: http://dx.doi.org/10.1007/978-3-030-87240-3_31 (citado nas pgs. 1, 15).

[Glorot e Bengio 2010] Xavier Glorot e Yoshua Bengio. "Understanding the difficulty of training deep feedforward neural networks". In: *Proceedings of the Thirteenth International Conference on Artificial Intelligence and Statistics*. Ed. por Yee Whye Teh e Mike Titterington. Vol. 9. Proceedings of Machine Learning Research. Chia Laguna Resort, Sardinia, Italy: PMLR, 2010, pp. 249–256. url: https://proceedings.mlr.press/v9/glorot10a.html (citado na pg. 13).

[Goodfellow *et al.* 2016] Ian Goodfellow, Yoshua Bengio e Aaron Courville. *Deep Learning*. http://www.deeplearningbook.org. MIT Press, 2016 (citado nas pgs. 14, 15).

[Grzybowski e Brona 2023] A. Grzybowski e P. Brona. "Approval and certification of ophthalmic ai devices in the european union". *Ophthalmology and Therapy* 12.2 (2023), pp. 633–638. doi: 10.1007/s40123-023-00652-w (citado na pg. 2).

[Kaiming He *et al.* 2016] Kaiming He, Xiangyu Zhang, Shaoqing Ren e Jian Sun. "Deep residual learning for image recognition". In: *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. 2016, pp. 770–778 (citado na pg. 18).

[Kelei He *et al.* 2023] Kelei He *et al.* "Transformers in medical image analysis". *Intelligent Medicine* 3.1 (2023), pp. 59–78. issn: 2667-1026. doi: https://doi.org/10.1016/j.imed.2022.07.002. url: https://www.sciencedirect.com/science/article/pii/S2667102622000717 (citado na pg. 22).

[Hernández, Smith *et al.* 2023] Jaime Hernández, Carol Smith *et al.* "Example title for pmc article". *Journal of Example Medicine* 10.4 (2023), pp. 123–130. doi: 10.1234/example.doi. url: https://pmc.ncbi.nlm.nih.gov/articles/PMC10436766/ (citado na pg. 3).

[Huang *et al.* 2017] Gao Huang, Zhuang Liu, Laurens Van Der Maaten e Kilian Q Weinberger. "Densely connected convolutional networks". In: *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. 2017, pp. 4700–4708 (citado na pg. 19).

[International Diabetes Federation 2021] International Diabetes Federation. *IDF Diabetes Atlas*. 10th. 2021. url: https://diabetesatlas.org/ (citado na pg. 9).

[Ioffe e Szegedy 2015] Sergey Ioffe e Christian Szegedy. "Batch normalization: accelerating deep network training by reducing internal covariate shift". *CoRR* abs/1502.03167 (2015). arXiv: 1502.03167. url: http://arxiv.org/abs/1502.03167 (citado na pg. 14).

[Jabbar *et al.* 2022] M.K. Jabbar, J. Yan, H. Xu, Z. Ur Rehman e A. Jabbar. "Transfer learning-based model for diabetic retinopathy diagnosis using retinal images". *Brain Sciences* 12.535 (2022). doi: 10.3390/brainsci12040535 (citado nas pgs. 21, 23).

[Karthik *et al.* 2019] Karthik, Maggie e Sohier Dane. *APTOS 2019 Blindness Detection*. https://kaggle.com/competitions/aptos2019-blindness-detection. Kaggle. 2019 (citado na pg. 22).

[Kaushik *et al.* 2021] H. Kaushik *et al.* "Diabetic retinopathy diagnosis from fundus images using stacked generalization of deep models". *IEEE Access* 9 (2021), pp. 118121–118133. doi: 10.1109/ACCESS.2021.xxxxxx (citado na pg. 23).

[Kingma e Ba 2015] Diederik P. Kingma e Jimmy Ba. "Adam: A method for stochastic optimization". In: *3rd International Conference on Learning Representations, ICLR 2015, San Diego, CA, USA, May 7-9, 2015, Conference Track Proceedings*. Ed. por Yoshua Bengio e Yann LeCun. 2015. url: http://arxiv.org/abs/1412.6980 (citado na pg. 14).

[Krizhevsky *et al.* 2012] Alex Krizhevsky, Ilya Sutskever e Geoffrey E Hinton. "Imagenet classification with deep convolutional neural networks". In: *Advances in Neural Information Processing Systems (NIPS)*. 2012, pp. 1097–1105 (citado nas pgs. 1, 11, 17).

[LeCun *et al.* 1998] Yann LeCun, Léon Bottou, Yoshua Bengio e Patrick Haffner. "Gradient-based learning applied to document recognition". *Proceedings of the IEEE* 86.11 (1998), pp. 2278–2324 (citado nas pgs. 1, 11, 14, 17).

[Li *et al.* 2019] Tao Li *et al.* "Diagnostic assessment of deep learning algorithms for diabetic retinopathy screening". *Information Sciences* (2019). doi: 10.1016/j.ins.2019.06.011 (citado na pg. 1).

[Liu *et al.* 2021] Ze Liu *et al.* "Swin transformer: hierarchical vision transformer using shifted windows". In: *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*. 2021, pp. 10012–10022. url: https://arxiv.org/abs/2103.14030 (citado na pg. 19).

[Martinez-Murcia *et al.* 2021] Francisco J. Martinez-Murcia, Andrés Ortiz, Javier Ramírez, Juan M. Górriz e Ricardo Cruz. "Deep residual transfer learning for automatic diagnosis and grading of diabetic retinopathy". *Neurocomputing* 452 (2021), pp. 424–434. doi: 10.1016/j.neucom.2020.04.148. url: https://doi.org/10.1016/j.neucom.2020.04.148 (citado nas pgs. 1, 21).

[Mutawa e Sruthi 2022] A. M. Mutawa e Sai Sruthi. "Diabetic retinopathy classification using vision transformer". In: *2022 6th European Conference on Electrical Engineering & Computer Science (ELECS)*. IEEE, 2022, pp. 25–30. doi: 10.1109/ELECS55825.2022.00012. url: https://ieeexplore.ieee.org/document/10144062 (citado na pg. 22).

[Nair e Hinton 2010] Vinod Nair e Geoffrey E Hinton. "Rectified linear units improve restricted boltzmann machines". In: *ICML 2010*. 2010, pp. 807–814 (citado na pg. 13).

[L. F. Nakayama *et al.* 2023] L. F. Nakayama *et al. A Brazilian Multilabel Ophthalmological Dataset (BRSET) (version 1.0.0)*. 2023. doi: 10.13026/xcxw-8198 (citado nas pgs. 1, 16).

[Luis Filipe Nakayama *et al.* 2024] Luis Filipe Nakayama *et al.* "Brset: a brazilian multilabel ophthalmological dataset of retina fundus photos". *PLOS Digital Health* 3 (jul. de 2024), pp. 1–16. doi: 10.1371/journal.pdig.0000454. url: https://doi.org/10.1371/journal.pdig.0000454 (citado na pg. 26).

[Nazih *et al.* 2023] Waleed Nazih, Ahmad O. Aseeri, Osama Youssef Atallah e Shaker El-Sappagh. "Vision transformer model for predicting the severity of diabetic retinopathy in fundus photography-based retina images". *IEEE Access* 11 (2023), pp. 117546–117559. doi: 10.1109/ACCESS.2023.3326528. url: https://ieeexplore.ieee.org/document/10290868 (citado na pg. 22).

[Neimark 2024] Jill Neimark. *Retinal Hemorrhage*. https://www.allaboutvision.com/conditions/retinal-hemorrhage/. (Acesso em 20/12/2024) (citado na pg. 4).

[Ophthalmology AAO 2016] American Academy of Ophthalmology AAO. *Diabetic Retinopathy*. Accessed: [Insert date]. 2016. URL: https://www.aao.org/education/topic-detail/diabetic-retinopathy-europe (citado nas pgs. 1, 2).

[Polyak 1964] Boris Polyak. "Some methods of speeding up the convergence of iteration methods". *Ussr Computational Mathematics and Mathematical Physics* 4 (dez. de 1964), pp. 1–17. DOI: 10.1016/0041-5553(64)90137-5 (citado na pg. 14).

[Porwa *et al.* 2020] Prasanna Porwa *et al.* "Idrid: diabetic retinopathy – segmentation and grading challenge". *Medical Image Analysis* (2020). DOI: 10.1016/j.media.2019.101561 (citado na pg. 1).

[Porwal *et al.* 2020] Prasanna Porwal *et al.* "Idrid: diabetic retinopathy – segmentation and grading challenge". *Medical Image Analysis* (2020). DOI: 10.1016/j.media.2019.101561 (citado na pg. 22).

[Sadeghzadeh *et al.* 2023] Arezoo Sadeghzadeh, Masum Shah Junayed, Tarkan Aydin e Md Baharul Islam. "Hybrid cnn+transformer for diabetic retinopathy recognition and grading". In: *2023 Innovations in Intelligent Systems and Applications Conference (ASYU)*. 2023, pp. 1–6. DOI: 10.1109/ASYU58738.2023.10296789 (citado na pg. 22).

[Saini *et al.* 2023] Naveen Kumar Saini, Debangana Ram e Manasi Gyanchandani. "Multi-headed cnn and vision transformer-based diabetic retinopathy classification". In: *2023 14th International Conference on Computing Communication and Networking Technologies (ICCCNT)*. 2023, pp. 1–6. DOI: 10.1109/ICCCNT56998.2023.10306806 (citado na pg. 22).

[Sebastian *et al.* 2023] Anila Sebastian, Omar Elharrouss, Somaya Al-Maadeed e Noor Almaadeed. "A survey on deep-learning-based diabetic retinopathy classification". *Diagnostics* 13.345 (2023). Accessed: [Insert Date]. DOI: 10.3390/diagnostics13030345. URL: https://doi.org/10.3390/diagnostics13030345 (citado nas pgs. 1, 21).

[Shang *et al.* 2024] Fangxin Shang *et al. SynFundus-1M: A High-quality Million-scale Synthetic fundus images Dataset with Fifteen Types of Annotation*. 2024. arXiv: 2312.00377 [cs.CV]. URL: https://arxiv.org/abs/2312.00377 (citado na pg. 27).

[Simonyan e Zisserman 2014] Karen Simonyan e Andrew Zisserman. "Very deep convolutional networks for large-scale image recognition". *arXiv preprint arXiv:1409.1556* (2014) (citado na pg. 18).

[Srivastava *et al.* 2014] Nitish Srivastava, Geoffrey Hinton, Alex Krizhevsky, Ilya Sutskever e Ruslan Salakhutdinov. "Dropout: a simple way to prevent neural networks from overfitting". *Journal of Machine Learning Research* 15.56 (2014), pp. 1929–1958. url: http://jmlr.org/papers/v15/srivastava14a.html (citado na pg. 14).

[Suma e Kumar 2018] K.G. Suma e V.S. Kumar. "A quantitative analysis of histogram equalization-based methods on fundus images for diabetic retinopathy detection". *In Computational Intelligence and Big Data Analytics* (2018), pp. 55–63 (citado na pg. 23).

[Szegedy *et al.* 2015] Christian Szegedy *et al.* "Going deeper with convolutions". In: *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. 2015, pp. 1–9 (citado na pg. 18).

[Teo *et al.* 2021] Z. L. Teo, Y.-C. Tham, M. Yu *et al.* "Global prevalence of diabetic retinopathy and projection of burden through 2045: systematic review and meta-analysis". *Ophthalmology* 128.11 (2021), pp. 1580–1591. url: https://www.aaojournal.org/article/S0161-6420%2821%2900321-3/fulltext (citado nas pgs. 1, 9).

[Wu *et al.* 2017] Bo Wu, Weifang Zhu, Fei Shi, Shuxia Zhu e Xinjian Chen. "Automatic detection of microaneurysms in retinal fundus images". *Computerized Medical Imaging and Graphics* 55 (2017). Special Issue on Ophthalmic Medical Image Analysis, pp. 106–112. issn: 0895-6111. doi: https://doi.org/10.1016/j.compmedimag.2016.08.001. url: https://www.sciencedirect.com/science/article/pii/S0895611116300787 (citado na pg. 4).

[Yaqoob *et al.* 2021] Muhammad Kashif Yaqoob, Syed Farooq Ali, Muhammad Bilal, Muhammad Shehzad Hanif e Ubaid M. Al-Saggaf. "Resnet based deep features and random forest classifier for diabetic retinopathy detection". *Sensors* 21.11 (2021), p. 3883. doi: 10.3390/s21113883. url: https://www.mdpi.com/article/10.3390/s21113883 (citado na pg. 21).

[Zhang *et al.* 2018] Hongyi Zhang, Moustapha Cisse, Yann N. Dauphin e David Lopez-Paz. *mixup: Beyond Empirical Risk Minimization*. 2018. arXiv: 1710.09412 [cs.LG]. url: https://arxiv.org/abs/1710.09412 (citado na pg. 15).