RELATÓRIO

CLASSIFICAÇÃO DE GLAUCOMA EM IMAGENS DE FUNDO DE OLHO COM APRENDIZAGEM PROFUNDA

Leandro Zangirolami Trovões (Orientando)

Carlos da Silva dos Santos (Orientador)

Universidade Federal do ABC

Santo André,

Abril de 2025

Resumo

O glaucoma é a principal causa de cegueira irreversível no mundo, com projeção de atingir até 111,8 milhões de pessoas em 2040. Caracterizada pelo dano progressivo ao nervo óptico, a doença é assintomática em seus estágios iniciais, tornando o diagnóstico precoce essencial para seu controle. O exame de fundo de olho permite a identificação de sinais característicos da doença, e técnicas de aprendizagem profunda têm demonstrado grande potencial na triagem automatizada.

Este trabalho propôs o desenvolvimento de um sistema baseado em redes neurais convolucionais para a detecção de casos referenciáveis para avaliação de glaucoma em imagens de fundo de olho. O sistema é composto por três etapas principais: segmentação do disco óptico, classificação binária para glaucoma, e classificação multi-rótulo das características clínicas justificadoras, utilizando o conjunto de dados JustRAIGS.

No conjunto de teste definido, o modelo atingiu sensitividade de 89,72% em 95% de especificidade e AUC-ROC de 0,9781 na tarefa de classificação binária. Na tarefa de classificação multi-rótulo, obteve distância de Hamming de 0,12960. O sistema apresentou desempenho competitivo em relação a trabalhos recentes da literatura, demonstrando sua aplicabilidade como ferramenta auxiliar na triagem de glaucoma.

Santo André, abril de 2025

Abstract

Glaucoma is the leading cause of irreversible blindness worldwide, with projections estimating up to 111.8 million affected individuals by 2040. Characterized by progressive damage to the optic nerve, the disease is asymptomatic in its early stages, making early diagnosis crucial for its management. Fundus imaging enables the identification of characteristic signs of glaucoma, and deep learning techniques have shown great potential for automated screening.

This work proposes the development of a system based on convolutional neural networks for the detection of referable glaucoma cases in fundus images. The system comprises three main stages: optic disc segmentation, binary classification for glaucoma, and multi-label classification of clinical justifications, using the JustRAIGS dataset.

On the defined test set, the proposed model achieved a sensitivity of 89.72% at 95% specificity and an AUC-ROC of 0.9781 in the binary classification task. In the multi-label classification task, it achieved a Hamming distance of 0.12960. The system demonstrated competitive performance compared to recent works in the literature, supporting its applicability as an auxiliary tool for glaucoma screening.

Santo André, abril de 2025

Conteúdo

1 Introdução

Principal causa de cegueira irreversível no mundo [31], o glaucoma é uma doença sem cura, caracterizada pelo dano progressivo ao nervo óptico [37]. Estima-se que, em 2020, 3,6 milhões de pessoas com 50 anos ou mais já tenham perdido a visão para o Glaucoma [31] e um estudo de 2014 ainda projeta que 111.8 milhões de pessoas sejam afetadas pela doença no ano de 2040 [34].

O estágio inicial do glaucoma é assintomático, mas a doença causa perda progressiva da visão periférica conforme avança e pode levar à perda total da visão. O tratamento pode retardar ou prevenir a progressão, mas depende de um diagnóstico precoce, geralmente antes mesmo dos primeiros sintomas [37].

Uma das formas de identificar a presença de glaucoma é por meio da *fundoscopia* ou exame de fundo de olho, no qual é possível observar alterações características, muitas vezes antes mesmo que a perda de visão se torne detectável [35]. Um exemplo de imagem obtida nesse exame é apresentado na Figura 1.

A principal característica observada ao analisar o fundo de olho é o tamanho da *escavação* em relação ao tamanho do disco óptico do paciente, ambos destacados na Figura 2. O disco óptico é a região em que as células da retina se convergem para formar o nervo óptico. Essa convergência forma uma depressão ao centro do disco, chamada de escavação (em inglês conhecido como *optic cup*). O *anel neurorretiniano* é a região que envolve a escavação. A razão entre o tamanho do disco e da escavação é conhecida como razão disco-copo (em inglês *cup-to-disk ratio*) e seu valor acima do normal é um indicativo do dano causado pela doença [35].

Figura 1: Imagem de fundo de olho. Obtida do banco de dados JustRAIGS [22].

Com o avanço das técnicas de aprendizagem profunda, modelos de redes neurais

convolucionais têm se mostrado promissores na análise automática de imagens de fundo de olho para auxílio no diagnóstico do glaucoma [16], sendo capazes de até mesmo superar a avaliação humana [33]. No entanto, desafios persistem, como a escassez de grandes bases de dados anotadas e a necessidade de aumentar a interpretabilidade dos modelos [16].

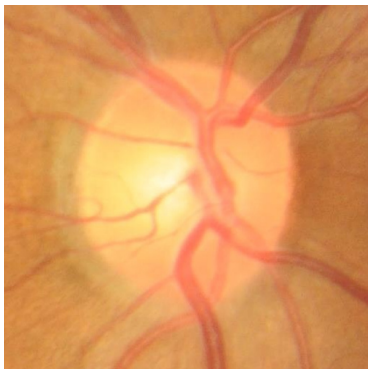

Figura 2: Disco óptico em destaque. Obtida do banco de dados JustRAIGS [22].

Neste trabalho, exploramos o uso de aprendizagem profunda para a detecção de casos referenciáveis de glaucoma em imagens de fundo de olho. Para isso, utilizamos o banco de dados JustRAIGS [22], que além das classificações principais, inclui anotações adicionais que justificam a decisão clínica.

O conjunto JustRAIGS foi disponibilizado como parte da competição promovida pelo 21º *IEEE International Symposium on Biomedical Imaging* (ISBI 2024), intitulada *Justified Referral in AI Glaucoma Screening — JustRAIGS*. Ela é composta por duas tarefas principais: (i) predizer se a imagem de fundo de olho deve ser referenciada para avaliação de glaucoma e (ii) predizer quais indicativos de diagnóstico justificam a referência. O sistema desenvolvido neste trabalho adota a mesma estrutura proposta.

Como a competição já havia se encerrado no momento da realização deste trabalho e o conjunto de avaliação é fechado, não é possível determinar qual classificação nosso modelo alcançaria. Entretanto, empregamos um conjunto de teste independente, o que possibilita uma comparação aproximada com os resultados publicados. Nesse conjunto, nosso modelo obteve uma sensitividade de 89,72% em 95% de especificidade mínima na primeira tarefa e uma distância de Hamming de 0,12960 na segunda. Para referência, durante a competição, Zhang et al. [38] alcançaram uma sensitividade de 87,00% em 95% de especificidade e distância de Hamming de 0,239, enquanto Kubrak [13] obteve 90,90% de sensitividade e distância de Hamming de 0,1280.

O restante deste texto está organizado da seguinte maneira: a Seção 2 apresenta

os objetivos do trabalho. A Seção 3 discute conceitos fundamentais de aprendizagem profunda e visão computacional aplicados à tarefa, além de apresentar os bancos de dados utilizados e trabalhos relacionados. A metodologia proposta é detalhada na Seção 4, os resultados obtidos são discutidos na Seção 5, e as conclusões são apresentadas na Seção 6.

2 Objetivos

O objetivo geral deste trabalho é desenvolver e avaliar um sistema baseado em aprendizagem profunda para a identificação de casos referenciáveis de glaucoma em imagens de fundo de olho, buscando também fornecer predições sobre características adicionais (p.ex. presença de hemorragia) que justifiquem a decisão do modelo.

Os objetivos específicos são:

- Desenvolver um método de segmentação do disco óptico, delimitando a região de interesse (ROI) nas imagens de fundo de olho;
- Treinar um classificador binário para glaucoma;
- Treinar um classificador multi-rótulo para identificar características específicas associadas ao glaucoma, com base nas anotações adicionais do conjunto JustRAIGS;
- Comparar os resultados obtidos com aqueles reportados na literatura recente, especialmente trabalhos utilizando o conjunto de dados JustRAIGS.

3 Revisão Bibliográfica

Iniciamos esta seção com uma revisão dos fundamentos de aprendizagem profunda e redes neurais convolucionais. Em seguida, apresentamos algumas bases de dados públicas utilizadas na construção de sistemas automatizados para detecção de glaucoma, bem como trabalhos anteriores relacionados. Detalhamos o conjunto de dados JustRAIGS e a competição a ele atrelado, que serve de referência para este trabalho, e apresentamos alguns dos trabalhos submetidos. Por fim, descrevemos as principais métricas empregadas na avaliação de classificadores e segmentadores.

3.1 Aprendizado de Máquina

O aprendizado de máquina é uma subárea da inteligência artificial, definida em 1959 por Arthur Samuel como o campo de estudo que dá aos computadores a habilidade de aprender sem serem explicitamente programados. Posteriormente, Tom Mitchell (1997) formalizou o conceito, definindo que um programa de computador **aprende** com a experiência E em relação a uma classe de tarefas T e uma medida de desempenho P, se seu desempenho nas tarefas de T, medido por P, melhora com a experiência E [21].

As tarefas de aprendizado de máquina podem ser classificadas em diferentes categorias, de acordo com o tipo de sinal de resposta (*feedback*) disponível para o sistema [27]:

- **Aprendizado não supervisionado**: o algoritmo deve identificar padrões nos dados de entrada sem receber *feedback* explícito. A tarefa mais comum nessa categoria é o agrupamento de dados (*clustering*);
- **Aprendizado supervisionado**: o algoritmo recebe pares de dados entrada-saída e deve aprender uma função que mapeie as entradas para as saídas. Um dos principais problemas nesse contexto é a classificação, em que o objetivo é atribuir um rótulo a novas entradas com base no conhecimento adquirido durante o treinamento;
- **Aprendizado semi-supervisionado**: ocorre quando apenas parte dos dados está rotulada, combinando técnicas de aprendizado supervisionado e não supervisionado para explorar a estrutura dos dados;
- **Aprendizado por reforço**: um agente aprende a tomar decisões por meio de interações com um ambiente, recebendo recompensas ou penalidades a partir de suas ações, de modo a maximizar o retorno esperado.

3.2 Redes Neurais e Aprendizado Profundo

Redes neurais artificiais são modelos computacionais inspirados na estrutura do cérebro humano, compostos por unidades chamadas neurônios artificiais, geralmente organizados em camadas. Cada neurônio realiza uma operação matemática simples, mas a composição de múltiplas camadas permite ao modelo aprender representações complexas dos dados [8].

O *aprendizado profundo* (*deep learning*) é uma subdivisão do aprendizado de máquina que utiliza redes neurais com múltiplas camadas ocultas. Essas camadas permitem que o modelo aprenda representações hierárquicas de características a partir de dados de entrada [8].

O avanço no treinamento de redes profundas, impulsionado por maiores volumes de dados, poder computacional e inovações algorítmicas como o uso de funções de ativação não-lineares e regularização, levou ao ressurgimento do interesse por redes neurais na última década, consolidando o aprendizado profundo como a principal abordagem em diversas áreas de inteligência artificial [8].

3.3 Redes Neurais Convolucionais

As redes neurais convolucionais (*Convolutional Neural Networks* - CNNs) são uma arquitetura especializada de redes neurais profundas, proposta inicialmente para lidar eficientemente com dados que apresentam estrutura espacial bidimensional, como imagens [14]. São caracterizadas pela aplicação de operações de convolução em camadas sucessivas, permitindo que o modelo aprenda padrões locais, como bordas e texturas, e, progressivamente, composições mais complexas, como formas e objetos inteiros.

Uma CNN é composta pela repetição de camadas convolucionais, comumente intercaladas com camadas de subamostragem (*pooling*) e, ao final, camadas totalmente conectadas (*fully connected*). As camadas convolucionais extraem representações locais dos dados de entrada por meio de filtros deslizantes aprendidos, enquanto as camadas de subamostragem reduzem a dimensionalidade espacial, aumentando a robustez a pequenas variações e deslocamentos [8]. As camadas totalmente conectadas realizam a combinação final das características extraídas para tarefas como classificação — que consiste em atribuir um rótulo a uma entrada — ou regressão, em que o objetivo é prever um valor numérico contínuo.

Apesar de conceitos precursores das CNNs terem surgido na década de 1980, e do primeiro modelo prático ter sido apresentado por LeCun et al. na década de 1990 [14], as CNNs só ganharam ampla popularidade com o surgimento do modelo AlexNet em 2012, que revolucionou a área ao vencer o desafio *ImageNet Large Scale Visual Recognition Challenge* (ILSVRC) [12]. A AlexNet demonstrou que, com volume de dados e poder computacional suficientes, aliados ao uso de técnicas de regularização — que visam reduzir

o sobreajuste do modelo aos dados de treinamento, como o *dropout* — e funções de ativação não lineares — como a ReLU, que introduzem não linearidade nas camadas —, era possível treinar redes convolucionais profundas com desempenho superior aos modelos tradicionais de aprendizado de máquina.

Após esse avanço, diversas novas arquiteturas de CNNs foram propostas, como a VGG [28] (2014), a GoogLeNet (posteriormente conhecida como Inception) [32] (2014), a ResNet [9] (2015) e a MobileNet [10] (2017). Não demorou muito para as CNNs se popularizaram como uma metodologia de escolha para a análise de imagens médicas, com aplicações em classificação, detecção de objetos, segmentação e outros [18]. Em particular, seu uso em imagens de fundo de olho para diagnóstico de doenças oculares, incluindo a identificação de glaucoma, também ganhou destaque nos últimos anos [16].

A ResNet, proposta por He et al. em 2015 [9], introduziu o conceito de conexões residuais, que permitem o treinamento de redes muito profundas, mitigando o problema da degradação do desempenho. Essas conexões consistem em atalhos que somam a entrada de uma camada diretamente à sua saída, facilitando o fluxo do gradiente durante o treinamento. A arquitetura ResNet obteve resultados de destaque, vencendo o ILSVRC de 2015. Ela se tornou uma das arquiteturas mais populares e é inclusive uma das mais utilizadas na análise de imagens de fundo de olho [16].

3.4 Transferência de Aprendizagem

A técnica de transferência de aprendizagem consiste em aproveitar os pesos aprendidos por um modelo em uma tarefa anterior para melhorar seu desempenho em uma nova tarefa. Ela é comumente aplicada quando a segunda tarefa possui poucos dados disponíveis para treinamento, como ocorre com frequência no contexto de imagens médicas [20].

O ImageNet [26] é um conjunto de dados extenso e diverso, publicamente disponível, composto por mais de 1 milhão de imagens anotadas em 1000 classes. Esse conjunto é comumente usado no pré-treinamento de modelos, já que eles aprendem representações visuais genéricas que podem ser reaproveitadas em tarefas específicas por meio da transferência de aprendizagem [20].

3.5 Bases de dados para identificação de glaucoma

O desenvolvimento de modelos automáticos para detecção de glaucoma depende da disponibilidade de bases de dados de imagens de fundo de olho rotuladas por especialistas. Diversos conjuntos já foram produzidos e parte deles disponibilizados nos últimos anos, permitindo avanços na área.

Entre os principais bancos de dados, destacamos:

- **ORIGA** [39]: lançado em 2010, contém 650 imagens coletadas entre 2004 e 2007 pelo *Singapore Eye Research Institute*, com anotações sobre a presença de glaucoma, medidas do disco óptico e da escavação, a identificação de hemorragias no disco, entre outros;

- **DRISHTI-GS1** [29] [30]: coletado e anotado pelo *Aravind Eye Hospital* em Madurai, Índia, foi disponibilizado em 2015 e inclui 101 imagens com rótulo de glaucoma e segmentações manuais do disco óptico e da escavação;

- **RIGA** [1]: introduzido em 2018, reúne 750 imagens provenientes de três diferentes fontes, com rótulos de glaucoma e anotações dos limites do disco óptico e da escavação feitas por oftalmologistas experientes.

- **ACRIMA** [4]: lançado como parte do programa ACRIMA, financiado pelo *Ministerio de Economía y Competitividad* da Espanha, contém 705 imagens rotuladas quanto à presença de glaucoma por dois especialistas;

- **RIM-ONE DL** [6]: lançado em 2020, conta com 485 imagens, todas rotuladas quanto à presença de glaucoma e com anotações de segmentação do disco óptico e da escavação;

- **REFUGE** [25]: contendo 1200 imagens, combina informações de diagnóstico de glaucoma e segmentações manuais do disco óptico;

- **JustRAIGS** [22]: lançado em 2024, é um dos maiores bancos disponíveis, com 101.442 imagens classificadas. Além da classificação principal (referenciável ou não para glaucoma), o JustRAIGS inclui dez anotações adicionais que registram sinais clínicos observados pelos avaliadores para justificar a escolha, tais como "Aparência

superior do anel neurorretiniano", "Exposição de vasos circunlineares inferior" ou "Hemorragias de disco".

Um resumo desses conjuntos de dados é apresentado na Tabela 1.

Nome	**Nº de imagens**	**Ano de publicação**
ORIGA	650	2010
DRISHTI-GS1	101	2015
RIGA	750	2018
ACRIMA	705	2019
RIM-ONE DL	485	2020
REFUGE	1.200	2020
JustRAIGS	101.442	2024

Tabela 1: Bancos de dados para avaliação de glaucoma.

3.6 Trabalhos relacionados

Diversos trabalhos recentes exploraram a aplicação de técnicas de aprendizagem de máquina para a detecção de glaucoma em imagens de fundo de olho, utilizando diferentes abordagens e bancos de dados. A Tabela 2 apresenta um quadro resumo dos trabalhos apresentados a seguir.

Noronha et al. (2014) utilizou cumulantes de alta ordem para identificar glaucoma em 272 imagens de fundo de olho obtidas por conta própria, obtendo acurácia de 84,72% [24]. Chen et al. (2015) implementou uma rede neural convolucional contendo seis camadas, sendo quatro convolucionais e 2 totalmente conectadas, a treinou com o ORIGA e o banco de dados privado SCES, obtendo de resultado os valores de área sobre curva ROC (AUC) 0,831 e 0,887 [3]. Liu et al. (2019) desenvolveu uma rede convolucional baseada na arquitetura ResNet e a treinou com 241.032 imagens, obtendo uma AUC de 0,996 [19]. Diaz-Pinto et al. (2019) compararam cinco modelos pré-treinados no ImageNet (VGG16, VGG19, InceptionV3, ResNet50 e Xception), utilizando cinco conjuntos de dados públicos, e constataram que o modelo Xception alcançou melhor desempenho com AUC 0,9605, especificidade 0,8580 e sensibilidade 0,9346.

Em seu estudo de revisão, a respeito do uso de aprendizagem profunda em imagens de fundo de olho, Li et al. (2021) [16] aponta como algumas das limitações e áreas para melhorias futuras a baixa disponibilidade de dados anotados com qualidade e a falta de interpretabilidade dos resultados, característica inerente à aprendizagem profunda.

Artigo	**Método**	**Banco de dados**	**ACC**	**AUC**
Noronha et al. (2014) [24]	Cumulantes	Privado com 272 imagens	0,847	-
Chen et al. (2015) [3]	CNN própria	ORIGA	-	0,831
Chen et al. (2015) [3]	CNN própria	ORIGA, SCES	-	0,887
Ferreira et al. (2018) [5]	U-Net	RIM-ONE, DRISHTI-GS, DRIONS-DB	1	1
Liu et al. (2019) [19]	ResNet	Privado com 241.032 imagens	0,996	-
Diaz-Pinto et al. (2019) [4]	Xception	ACRIMA, HRF, DRISHTI-GS1, RIM-ONE, sjchoi86-HRF	-	0,9605

Tabela 2: Trabalhos anteriores e resultados obtidos.

3.6.1 Competição JustRAIGS

Com a introdução do conjunto de dados JustRAIGS e da competição homônima promovida pelo 21º *IEEE International Symposium on Biomedical Imaging* (ISBI 2024), novos trabalhos foram desenvolvidos. A competição propõe o desenvolvimento e a avaliação de algoritmos de inteligência artificial para triagem de glaucoma em imagens de fundo de olho, utilizando o conjunto de dados JustRAIGS.

A competição é dividida em duas tarefas: (i) a classificação binária entre "referenciável para glaucoma" (*referable glaucoma* - RG) e "não referenciável para glaucoma" (*no referable glaucoma* - NRG), denominada *referral performance*; e (ii) a classificação multi-rótulo das dez características adicionais associadas ao glaucoma observáveis nas

imagens, denominada *justification performance*. Mais adiante iremos apresentas as anotações adicionais do JustRAIGS.

O conjunto de treinamento, disponibilizado publicamente, contém 101.442 imagens anotadas. O conjunto de teste, composto por 9.741 imagens, será mantido privado e utilizado exclusivamente para a avaliação dos modelos submetidos.

A avaliação da tarefa de classificação binária foi baseada na sensitividade em um ponto de operação de 95% de especificidade. Já a avaliação da tarefa multi-rótulo utilizou uma distância de Hamming modificada, que desconsidera rótulos nos quais os avaliadores humanos não concordaram. Quanto menor o valor da distância de Hamming, maior a concordância do modelo com os avaliadores humanos. Mais adiante iremos retomar a discussão sobre estas métricas.

Diversos trabalhos recentes foram publicados em virtude da competição do ISBI 2024. A Tabela 3 mostra um resumo dos trabalhos apresentados a seguir.

Galdran e Ballester [7] utilizaram uma ResNet50 com uma estratégia própria de suavização de rótulos, atingindo no conjunto de validação uma sensitividade de 94,33% em 95% de especificidade e uma distância de Hamming de 0,1440 para as anotações adicionais.

Zhang et al. [38] realizaram testes preliminares com mais de vinte arquiteturas de redes convolucionais e Vision Transformers (ViTs) — uma classe de modelos originalmente proposta para tarefas de linguagem natural, mas que tem sido adaptada para visão computacional. Para a tarefa de classificação binária, os autores optaram por um *ensemble* de cinco ResNet50 e, para a tarefa multi-rótulo, por um *ensemble* formado por uma ResNet50 e dois Transformers. O termo *ensemble* refere-se à combinação de predições de múltiplos modelos com o objetivo de melhorar a robustez e o desempenho geral. No conjunto de teste da competição, alcançaram sensitividade de 87% em 95% de especificidade e distância de Hamming de 0,239.

Kubrak [13] aplicou a segmentação da região de interesse utilizando um modelo YOLOv8, seguida do treinamento de um Vision Transformer para cada tarefa. Na avaliação da fase de desenvolvimento da competição, o modelo atingiu sensitividade de 90,90% em 95% de especificidade e distância de Hamming de 0,1280. No conjunto de validação próprio, quando a segmentação era bem-sucedida, atingiu respectivamente 93,25% e 0,1414.

Lin et al. [17] também utilizaram YOLOv8 para segmentação e Vision Transfor-

mers para classificação. No conjunto de validação próprio o modelo atingiu sensitividade de 91,59% em 95% de especificidade e distância de Hamming de 0,1417. Na avaliação de desenvolvimento da competição, atingiu respectivamente 85,70% e 0,1250.

Autor(es)	**Arquitetura(s)**	**Estratégia adicional**	**Sen. (95% Esp.)**	**Dist. de Hamming**
Galdran e Ballester [7]	ResNet50	Suavização de rótulos	94,33%*	0,1440*
Zhang et al. [38]	Ensemble ResNet50 + Transformers	-	87,00%	0,239
Kubrak [13]	YOLOv8 + Vision Transformer	Segmentação do disco	90,90%	0,1280
Lin et al. [17]	YOLOv8 + Vision Transformer	Segmentação do disco	85,70%	0,1250

Tabela 3: Resumo de trabalhos utilizando o conjunto JustRAIGS. Valores marcados com asterisco (*) correspondem a métricas obtidas em subconjuntos de validação; os demais valores foram reportados no conjunto de teste da competição.

3.7 Métricas de Classificação

A forma mais simples de avaliar o desempenho de um modelo de classificação binária é por meio da acurácia, isto é, a proporção de predições corretas [2]. Contudo, em conjuntos desbalanceados, a acurácia pode ser uma métrica enganosa, e outras métricas tornam-se mais relevantes [8].

As predições de um classificador binário podem ser organizadas conforme a matriz de confusão, com as seguintes definições:

- **Verdadeiro Positivo (VP)**: instâncias positivas corretamente classificadas como positivas;
- **Falso Positivo (FP)**: instâncias negativas incorretamente classificadas como positivas;
- **Verdadeiro Negativo (VN)**: instâncias negativas corretamente classificadas como negativas;
- **Falso Negativo (FN)**: instâncias positivas incorretamente classificadas como negativas.

A partir desses valores, diversas métricas podem ser calculadas:

Acurácia — Proporção de predições corretas:

$$\text{Acurácia} = \frac{VP + VN}{VP + FP + VN + FN} \tag{1}$$

Embora amplamente utilizada, a acurácia pode ser enganosa em conjuntos de dados desbalanceados.

Precisão — Proporção de predições positivas corretas:

$$\text{Precisão} = \frac{VP}{VP + FP} \tag{2}$$

Mede, dentre todas as instâncias classificadas como positivas, quantas realmente são positivas.

Sensibilidade ou *Recall* — Proporção de positivos corretamente identificados:

$$\text{Recall} = \frac{VP}{VP + FN} \tag{3}$$

Indica a capacidade do modelo de identificar corretamente instâncias positivas.

Especificidade — Proporção de negativos corretamente identificados:

$$\text{Especificidade} = \frac{VN}{VN + FP} \tag{4}$$

Reflete a capacidade de identificar corretamente instâncias negativas.

F1-Score — Média harmônica entre precisão e *recall*:

$$\text{F1} = 2 \times \frac{\text{Precisão} \times \text{Recall}}{\text{Precisão} + \text{Recall}} \tag{5}$$

É útil em cenários em que é necessário balancear entre precisão e *recall*.

Curva Precisão-*Recall* Representa a relação entre a precisão e o *recall* para diferentes limiares de decisão. É especialmente relevante em conjuntos de dados desbalanceados.

Curva ROC — A curva ROC (*Receiver Operating Characteristic*) plota a taxa de verdadeiros positivos contra a taxa de falsos positivos para diferentes limiares de decisão. A área sob a curva (AUC) é uma métrica comum derivada da ROC.

Sensibilidade em 95% de Especificidade — Uma métrica específica para cenários clínicos de triagem, em que se busca atingir alta especificidade. Avalia a sensitividade do modelo em um ponto de operação em que a especificidade é de 95% ou superior.

Distância de Hamming — Em problemas de classificação multi-rótulo, a distância de Hamming (normalizada) mede a proporção de discordância entre rótulos preditos e rótulos verdadeiros:

$$\text{Distância de Hamming} = \frac{1}{n}\sum_{i=1}^{n} \mathbb{I}(y_i \neq \hat{y}_i) \tag{6}$$

onde $\mathbb{I}$ é a função indicadora que devolve 1 se $y_i \neq \hat{y}_i$ e 0 caso contrário, y_i é o rótulo verdadeiro e $\hat{y}_i$ é a predição.

Na competição JustRAIGS, foi utilizada uma distância de Hamming modificada, ignorando rótulos em que os avaliadores não entraram em acordo.

3.8 Métricas de Segmentação

Índice de Jaccard ou Intersection over Union (IoU) — O índice de Jaccard, também conhecido como *Intersection over Union* (IoU), é uma métrica amplamente utilizada para avaliar a similaridade entre duas regiões segmentadas. É definido como a razão entre a área de interseção e a área da união entre a predição e o rótulo verdadeiro:

$$\text{IoU} = \frac{|\text{Predição} \cap \text{Verdadeiro}|}{|\text{Predição} \cup \text{Verdadeiro}|} \tag{7}$$

onde $|\cdot|$ denota a área (número de pixels) da respectiva região. Um valor de IoU igual a 1 indica sobreposição perfeita entre a segmentação prevista e a verdadeira, enquanto valores mais baixos indicam sobreposição parcial ou ausente.

Outra métrica comumente utilizada é a média da precisão média (*mean Average Precision* — mAP).

mAP50 — Corresponde à média da precisão média considerando apenas predições com *Intersection over Union* (IoU) superior a 50%.

mAP50-95 — Corresponde à média das precisões médias em múltiplos limiares de IoU, variando de 50% a 95% em intervalos de 5%. Esta métrica fornece uma avaliação mais rigorosa e detalhada da qualidade da segmentação.

4 Materiais e Métodos

Nesta seção, descrevemos as características do conjunto de dados e os tratamentos nele feitos, os passos realizados para a construção de um segmentador da região de interesse e os resultados obtidos, e os procedimentos e técnicas adotados para o treinamento de um classificador binário para identificar glaucoma e de um classificador multi-rótulo para identificar as anotações adicionais do JustRAIGS.

4.1 Conjunto de dados

Foi escolhido o conjunto de dados JustRAIGS [22] que contém 101.442 imagens de fundo de olho publicamente disponíveis, derivado do conjunto de dados REGAIS — *Rotterdam EyePACS Glaucoma AI Screening* [15]. As imagens foram providas pela *EyePACS LLC* em Santa Cruz, Califórnia nos EUA enquanto as anotações foram providas pelo Instituto de Oftalmologia de Roterdã (*Rotterdam Ophthalmic Institute*) do *Rotterdam Eye Hospital* em Roterdã nos Países Baixos.

As imagens foram obtidas nos Estados Unidos, por meio de um programa de rastreio de retinopatia diabética conduzido pela empresa EyePACS. A coleta ocorreu em aproximadamente 500 centros de triagem distribuídos pelo país, utilizando uma ampla variedade de equipamentos. A base é composta por participantes de múltiplas origens étnicas, incluindo latino-americanos (52%), pessoas brancas (8%), de ascendência africana (6%), asiáticos (4%), do subcontinente indiano (3%), nativos americanos (1%), de etnia mista (1%) e etnia não especificada (25%). A média de idade dos participantes foi de 57,1 anos, com desvio padrão de 10,4 anos [15].

Cada imagem foi anotada por dois avaliadores (A1 e A2), selecionados aleatoriamente dentre um conjunto de avaliadores qualificados. Para cada imagem, o avaliador

deveria escolher uma classificação principal entre referenciável para glaucoma (*referable glaucoma* - RG), não referenciável (*no referable glaucoma* - NRG) ou ainda se não era possível avaliar a imagem (*ungradable* - U). Referenciável para glaucoma deveria ser escolhido se fossem esperadas perdas no campo de visão para o olho avaliado. Caso os dois avaliadores concordassem na classificação principal, essa classificação era dada como final, caso contrário, um terceiro avaliador (A3), especialista em glaucoma, anotava a imagem e essa anotação era dada como final [15].

Para os casos em que o avaliador julgasse como referenciável para glaucoma, ele poderia selecionar até 10 anotação adicionais, que representam características aparentes de um olho com glaucoma, de forma a justificar sua escolha [15].

As anotações adicionais são:

- **ANRS**: Aparência superior do anel neurorretiniano (*Appearance neuroretinal rim superiorly*);
- **ANRI**: Aparência inferior do anel neurorretiniano (*Appearance neuroretinal rim inferiorly*);
- **RNFLDS**: Defeito na camada de fibras nervosas da retina superior (*Retinal nerve fiber layer defect superiorly*);
- **RNFLDI**: Defeito na camada de fibras nervosas da retina inferior (*Retinal nerve fiber layer defect inferiorly*);
- **BCLVS**: Exposição de vasos circunlineares superior (*Baring circumlinear vessel superiorly*);
- **BCLVI**: Exposição de vasos circunlineares inferior (*Baring circumlinear vessel inferiorly*);
- **NVT**: Nasalização do tronco vascular (*Nasalisation of vessel trunk*);
- **DH**: Hemorragias do disco óptico (*Disc hemorrhages*);
- **LD**: Pontos laminares (*Laminar dots*);
- **LC**: Escavação aumentada (*Large cup*).

Discordâncias nessas anotações adicionais, contudo, não foram resolvidas. Isto significa que há casos em que, apesar de A1 e A2 terem concordado na classificação principal como glaucoma, eles podem ter selecionado anotações adicionais diferentes para

justificar sua escolha e, nesses casos, A3 não foi requisitado. Sendo assim, o conjunto de dados inclui a anotação principal final e as anotações principal e adicionais de cada avaliador. Imagens cuja classificação final foi U não foram incluídas no JustRAIGS.

Para obter um valor final às anotações adicionais, no escopo desta seção, adotamos a seguinte regra para a resolução de discordância:

- A1 e A2 concordam na anotação principal → para cada anotação adicional, valor final é 1 se A1 e A2 marcaram como 1, do contrário 0;
- A1 concorda com A3 → aplicamos a regra acima para os valores de A1 e A3;
- A2 concorda com A3 → aplicamos a regra para os valores de A2 e A3;
- todos discordam → usamos os valores de A3.

As análises que seguem são baseadas no resultado obtido por meio desta regra.

4.1.1 Análise

Das 101.442 imagens que o conjunto de dados alega possuir, 19 não foram encontradas, restando 101.423. O conjunto é bastante desbalanceado: dentre todas as imagens, apenas 3.270 receberam anotação final para glaucoma, o que representa aproximadamente 3,22% ou 1 em cada 31.

Das anotações adicionais, ANRI é a mais frequente, aparecendo em 69,08% dos casos referenciáveis para glaucoma (ou 2,23% do total), DH por sua vez é a menos frequente, aparecendo em apenas 2,45% (0,08%). É possível observar certas correlações entre as anotações adicionais, apresentadas na Figura 3 por meio de uma matriz de correlação.

4.1.2 Imagens

A aquisição das imagens foi feita em vários centros de triagem com câmeras diferentes [15]. É possível observar grande variação de brilho, contraste e nitidez nas imagens do conjunto de dados, como ilustrado na Figura 4.

Nota-se também a presença de artefatos de aquisição variados em parte das imagens, como mostrado na Figura 5.

4.1.3 Divisão entre treino e teste

O conjunto de dados foi dividido na proporção 80% para treinamento e 20% para teste. Para realizar essa divisão, o conjunto foi inicialmente dividido entre NRG e RG. Para a

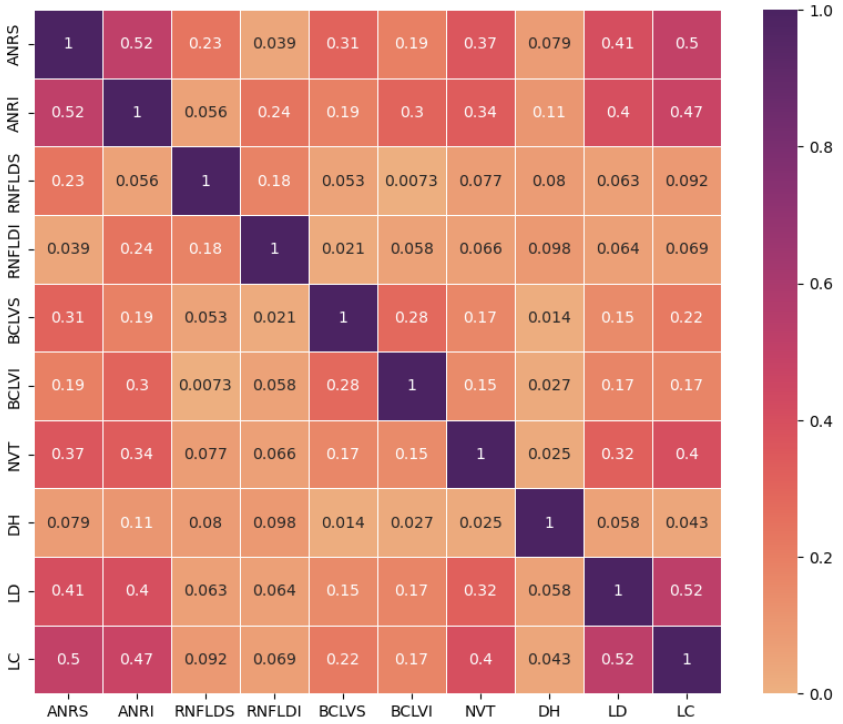

Figura 3: Matriz de correlação entre anotações adicionais.

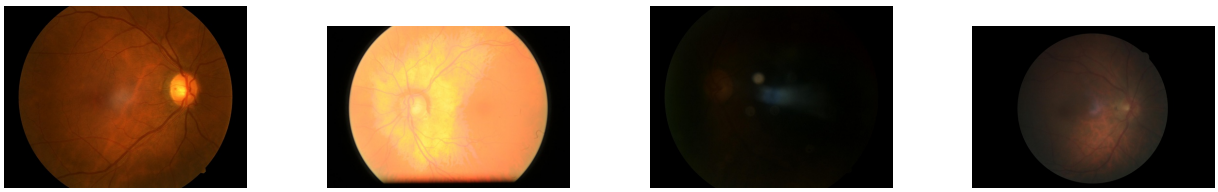

Figura 4: Variações de luminosidade e contraste entre imagens do JustRAIGS.

porção de NRG, foi feita uma divisão aleatória simples. Já para a porção de RG, com o intuito de preservar a proporção das anotações adicionais, foi realizada uma amostragem estratificada utilizando a classe *StratifiedShuffleSplit* e o método *split* da biblioteca *scikit-learn*.

A amostragem estratificada visa preservar, entre os conjuntos de treino e teste, a frequência relativa de cada classe. Como existem 10 anotações binárias, teoricamente seriam possíveis 2^{10} configurações diferentes. Na prática, no entanto, algumas combinações são muito mais frequentes, enquanto a imensa maioria sequer ocorre no conjunto de dados, de forma que não existem representantes suficientes para realizar a divisão nestes casos. Dessa forma, não foi possível estratificar diretamente considerando todas as anotações simultaneamente. Em vez disso, todas as possíveis permutações com 2, 3, 4 e 5 anotações

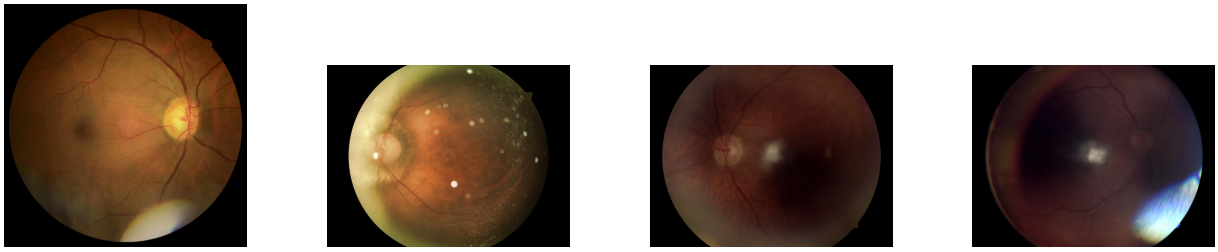

Figura 5: Imagens com artefatos de aquisição variados.

adicionais foram testadas. O uso de mais de 6 anotações se mostrou impraticável devido à escassez de representantes.

Para escolher a melhor combinação de anotações adicionais, foi utilizada uma métrica baseada na soma dos quadrados da diferença entre a proporção obtida e a desejada para cada anotação entre os dois conjuntos. Quanto menor a pontuação, mais balanceada e representativa era a divisão. Essa abordagem garantiu que a representatividade das anotações adicionais fosse preservada entre os conjuntos, mesmo para classes com baixa ocorrência.

4.2 Sistema proposto

O processo de detecção automatizada de glaucoma proposto neste trabalho envolve três etapas principais: (i) a segmentação do disco óptico na imagem de fundo de olho; (ii) a classificação da imagem como caso referenciável (RG) ou não referenciável (NRG), com base na região segmentada; e (iii) a identificação das razões que justificam a referência, por meio de um classificador multi-rótulo (*multi-label*), utilizando nas anotações adicionais presentes no JustRAIGS. A Figura 6 mostra um diagrama do sistema.

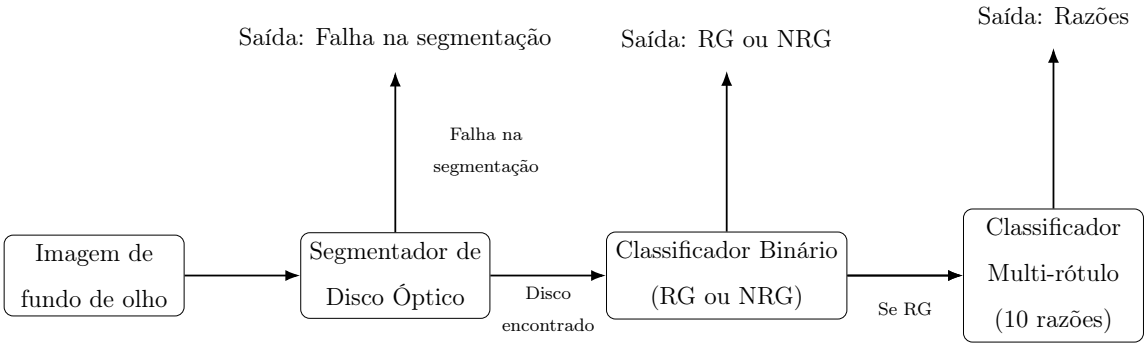

Figura 6: Diagrama do sistema proposto.

4.3 Segmentação da região de interesse

As características visuais associadas ao glaucoma, no exame de fundo de olho, estão ou no disco óptico ou em seu entorno [36]. Portanto, para melhor aproveitamento da rede neural classificadora, segmentamos as imagens na região do disco antes de utilizá-las em um classificador.

Para realizar essa segmentação, foi utilizado um modelo de aprendizagem profunda chamado Ultralytics YOLO11, capaz de realizar múltiplas tarefas de visão computacional em tempo real, como detecção de objetos, segmentação, classificação ou estimativa de pose humana. Ele pode ser treinado em novas bases de dados para aplicações específicas. Para cada tarefa existe uma família de modelos com quantidades de parâmetros diferentes [11]. Escolhemos o modelo *yolo11m* para detecção de objetos.

Para treinar o YOLO, foram aleatoriamente selecionadas 1400 imagens do conjunto de treinamento: 1000 para treino, 200 para validação e 200 para teste. Em cada um dos conjuntos a proporção de RG foi de 20%. Para cada imagem, o disco óptico foi manualmente anotado por meio de uma caixa delimitadora alinhadas aos eixos definida por dois pontos. As anotações foram feitas com o software AnyLabeling [23].

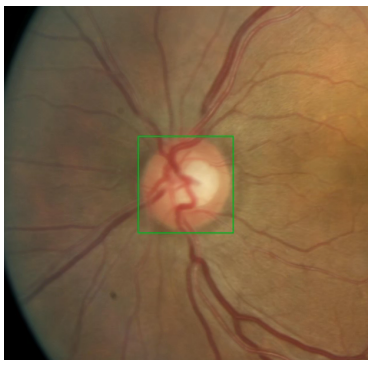

Figura 7: Disco óptico anotado no software AnyLabeling. Imagem original obtida do banco de dados JustRAIGS.

Foi feito o treinamento do segmentador com 100 épocas utilizando os parâmetros padrões do YOLO. Após o treinamento, no conjunto de validação o modelo atingiu precisão e *recall* 1, mAP50 0,995 e mAP50-95 0,945. No teste, apresentou precisão 0,995, *recall* 1, mAP50 0,995 e mAP50-95 0,936.

Em seguida, o segmentador foi aplicado para todo o conjunto de treinamento definido anteriormente na Seção 4.1.3. Das 81.138 imagens, em 80.725 (99,49%) o modelo

identificou um disco óptico, em 332 (0,41%) mais de uma detecção foi retornada e em 81 (0,10%) nenhuma. Para todas as detecções, utilizamos o *score* mínimo de 0,25, valor padrão do YOLO.

Analisando as 332 imagens para as quais o modelo retornou mais de uma predição, identificamos alguns artefatos de aquisição predominantes que confundiram o modelo. Algumas imagens com múltiplas detecções são apresentados na Figura 8.

Por outro lado, analisando as 81 imagens em que nenhuma detecção foi retornada, não conseguimos identificar uma causa predominante para a ausência de detecções. Listamos algumas das características dessas imagens, que se manifestam de formas variadas:

- disco óptico com limites pouco definidos;
- imagens com baixo contraste em que disco óptico aparece "apagado";
- imagens desfocadas ou com baixa nitidez;
- lesões ou deformações anatômicas diversas ao redor do disco óptico.

Alguns exemplos são apresentados na Figura 9.

Resolvemos então escolher aleatoriamente 100 imagens dentre as 332 com mais de uma detecção para treinar um novo segmentador. Foram anotadas manualmente e incluídas no conjunto de dados usado anteriormente, 50, 25 e 25 imagens para treino, validação e teste, respectivamente. Devido à falta de experiência médica para identificar corretamente o disco óptico em casos difíceis, adotamos uma postura cautelosa em não anotar as imagens para as quais o modelo não retornou nenhum resultado. Em uma aplicação real, casos como estes seriam encaminhados para avaliação de um médico. Um novo treinamento foi feito então, utilizando os mesmos parâmetros do anterior.

Desta vez, na validação o modelo atingiu precisão e *recall* 1, mAP50 0,995 e mAP50-95 0,944, enquanto que no teste atingiu precisão 0,999, *recall* 0,991, mAP50 0,995 e mAP50-95 0,934.

Ao aplicar esta segunda geração em todo o conjunto de treino, desta vez, o modelo identificou apenas um disco óptico em 81.016 imagens (99,85%), devolveu mais de uma detecção em 63 (0,08%) e nenhuma detecção em 59 (0,07%).

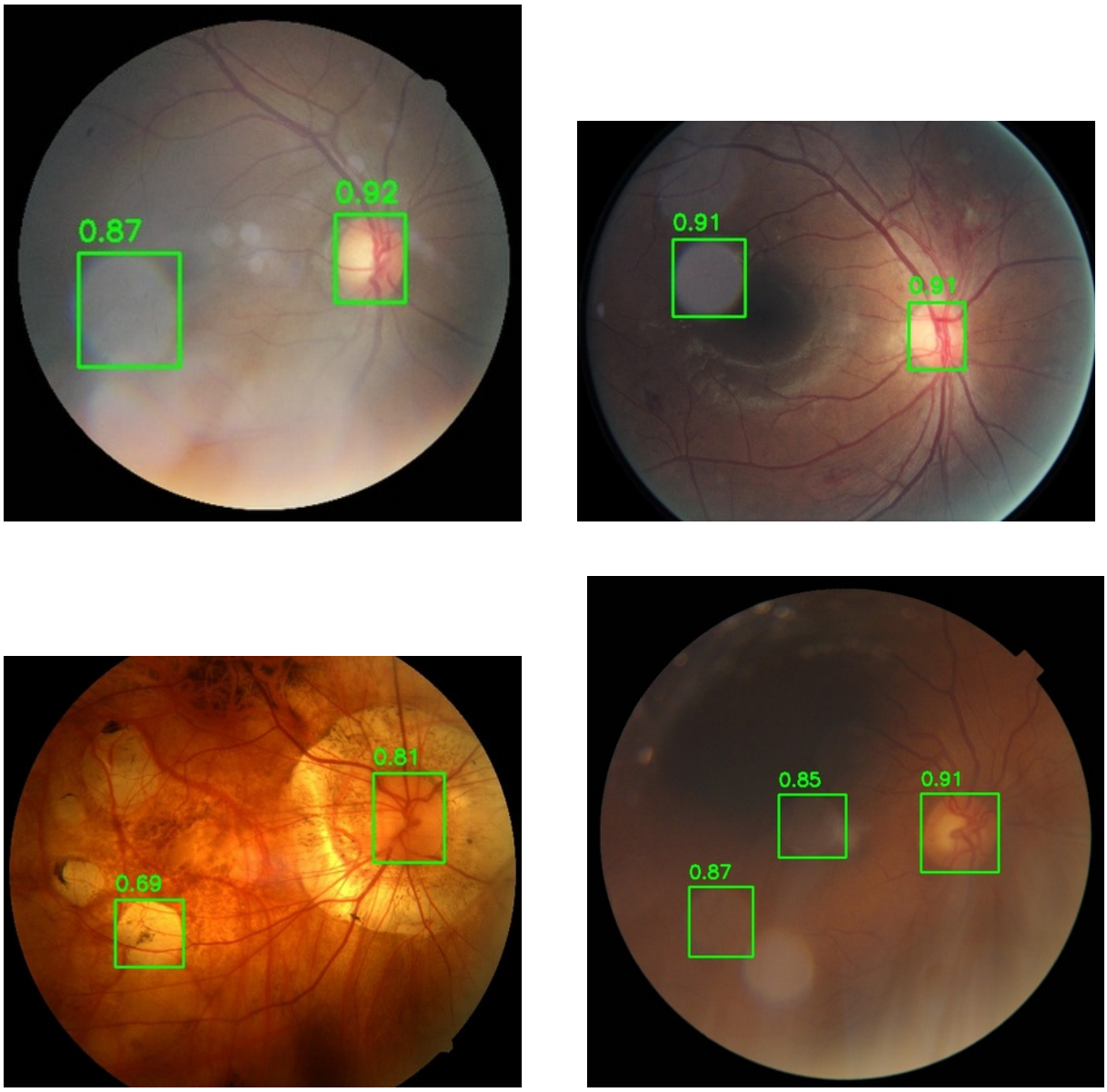

Figura 8: Imagens com múltiplas detecções pelo YOLO e *score* atribuído.

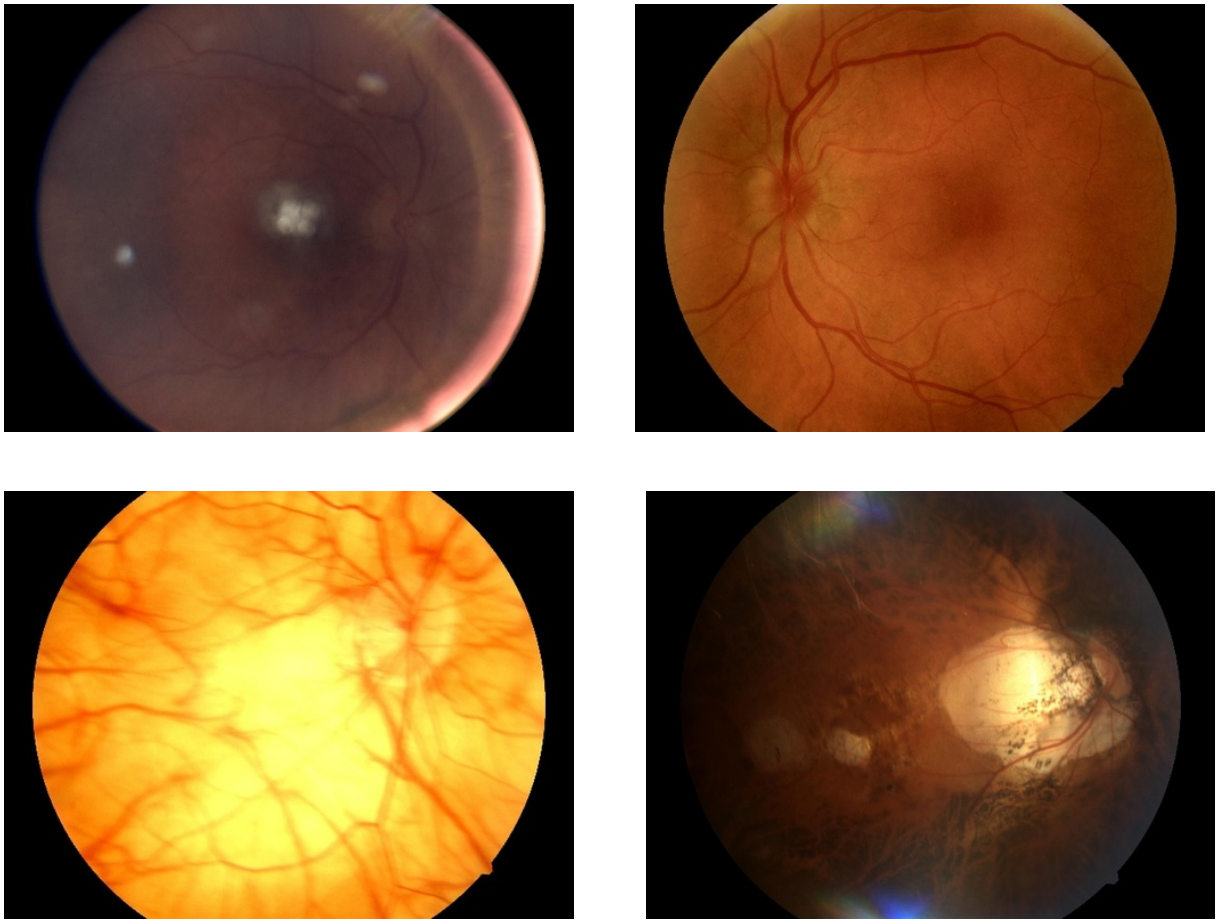

Figura 9: Imagens sem detecção pelo YOLO.

4.4 Classificação binária

Com as regiões de interesse já extraídas, a etapa de classificação binária consistiu em treinar um modelo capaz de distinguir entre casos referenciáveis para glaucoma (RG) e não referenciáveis (NRG). Para isso, utilizamos uma rede neural convolucional moderna empregando transferência de aprendizagem a partir de um modelo pré-treinado, como descrito a seguir.

Partindo de um modelo ResNet50 pré treinado no conjunto de dados ImageNet, removemos a camada superior e adicionamos uma camada GlobalAveragePooling2D, duas camadas totalmente conectadas de tamanhos 1024 e 256 com função de ativação ReLU e uma camada totalmente conectada de saída de tamanho 1 com função sigmoide, própria para classificação binária. Entre as novas camadas, foi aplicado *dropout* com taxa de 50%. Também experimentamos com uma arquitetura mais simples, de apenas uma camada totalmente conectada de tamanho 512 antes da saída, e outra arquitetura mais complexa, com camadas de tamanho na sequência 4096, 2048, 1024, 256 e saída.

As imagens disponíveis no conjunto de treino definido na Seção 4.1.3 foram subdivididas entre treino e validação para a etapa de treinamento do classificador. A divisão foi realizada de forma estratificada, mantendo a proporção entre as classes, alocando 80% das imagens para o treinamento e 20% para a validação, conforme o processo descrito anteriormente.

Com o objetivo de contornar o desbalanceamento entre as duas classes, todas as 2.610 imagens RG disponíveis no conjunto de treinamento foram espelhadas horizontalmente e salvas como novas imagens. Além disso, apenas um subconjunto de 10 mil imagens NRG aleatoriamente selecionadas foi utilizado, resultando em uma proporção entre as duas classes de aproximadamente 1 RG para 2 NRG. Em números absolutos, foram no conjunto de treinamento 4176 (2088 x 2) imagens RG e 8000 imagens NRG, enquanto que para validação foram 522 e 2000 imagens de cada classe respectivamente. Como não aplicamos o espelhamento no conjunto de validação, a proporção foi de aproximadamente 1 RG para 4 NRG. Também experimentamos uma subamostragem de NRG mais agressiva, para que a proporção fosse de 1 para 3, mas essa configuração apresentou resultado inferior à anterior.

Antes de serem utilizadas no classificador, todas as imagens foram recortadas no entorno do disco óptico detectado pelo segmentador. A região do corte foi definida com as

coordenadas devolvidas pelo modelo, acrescidas de uma margem de 50% para cada eixo.

Durante o treinamento, aplicamos algumas técnicas de aumentação de dados nas imagens:

- **Translação**: desloca a imagem vertical e horizontalmente;
- **Rotação**: gira a imagem por um ângulo aleatório;
- **Zoom**: aproxima ou afasta a imagem;
- **Espelhamento horizontal**: espelha horizontalmente a imagem com 50% de chance;
- **Brilho**: aumenta ou diminui o brilho;
- **Contraste**: ajusta aleatoriamente o contraste;
- **Cor**: converte a imagem de RGB para HSV, ajusta aleatoriamente o canal de cor (hue) e converte de volta para RGB.

Inicialmente experimentamos apenas com as transformações de translação, rotação, zoom e espelhamento, chamadas de transformações geométricas, que apenas movimentam a imagem. Posteriormente, experimentamos com as transformações de brilho, cor e contraste, chamadas de transformações fotométricas, que alteram o valores dos pixels. A Figura 10 mostra alguns exemplos de imagens antes e depois de passarem pelas transformações.

Fizemos experimentos com dois processos de treinamento: em duas etapas e em etapa única. O processo de duas etapas consiste em dividir o treinamento em uma primeira fase em que os pesos do modelo original são congelados e somente os pesos da novas camadas superiores são ajustados e em uma segunda fase, conhecida como ajuste fino (*fine-tuning*), em que todos os pesos são descongelados e o modelo é ajustado por inteiro a uma taxa de aprendizagem inferior. O processo de etapa única é equivalente ao ajuste fino, o treinamento já se inicia ajustando todos os pesos da rede à uma taxa de aprendizagem baixa.

As etapas de treinamento fizeram uso de parada antecipada (*early stopping*), uma técnica de regularização que interrompe o treinamento ao detectar que não houve melhora no desempenho do modelo no conjunto de validação após determinada quantidade de épocas. Como métrica a ser monitorada usamos a sensibilidade em 95% de especificidade.

Utilizamos o otimizador *ADAM*, com taxa de aprendizagem de 10^{-4} na primeira etapa e 10^{-5} no ajuste fino. Nos experimentos com somente uma etapa, utilizamos a

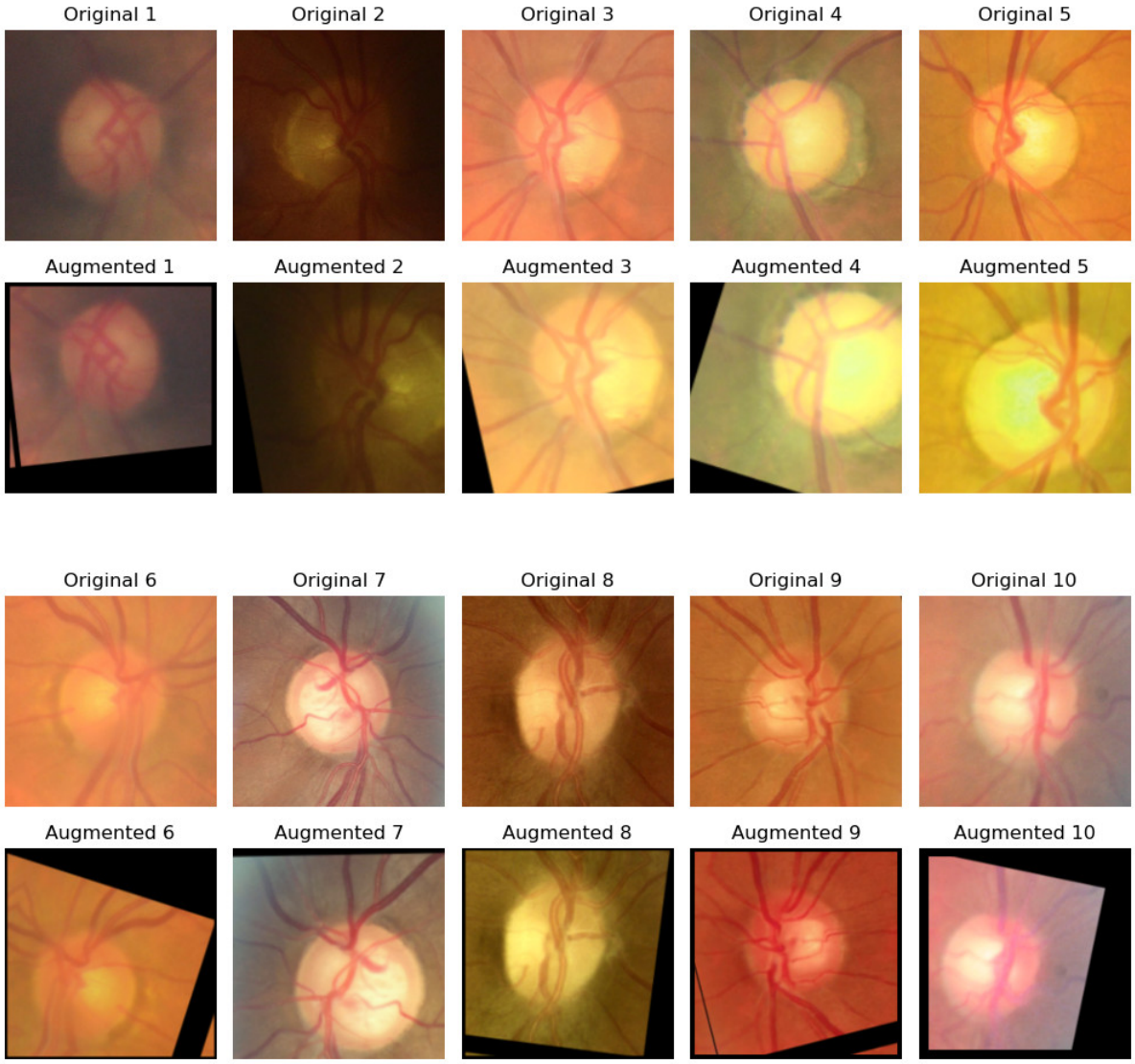

Figura 10: Imagens originais e após aplicação das aumentações de dados.

mesma taxa de aprendizagem do ajuste fino. Como função de perda utilizamos a entropia cruzada binária (*binary crossentropy*).

A Tabela 4 apresenta um quadro resumo das configurações de alguns dos experimentos que fizemos. O tamanho da rede se refere as camadas superiores adotadas na construção do modelo conforme descrito anteriormente, o processo se refere ao processo de treinamento em uma ou duas etapas e aumentações são as aumentações aplicadas nas imagens. Os modelos foram nomeados com base em suas configurações, utilizando as seguintes convenções: "P", "M" ou "G" para o tamanho da rede; "1" ou "2" para indicar a quantidade de etapas de treinamento; e o sufixo "f" quando há aplicação de transformação fotométrica.

Modelo	Tamanho da rede	Processo	Aumentações
M2	Médio	Duas etapas	Geométricas
G2	Grande	Duas etapas	Geométricas
M2f	Médio	Duas etapas	Geométricas e fotométricas
M1f	Médio	Etapa única	Geométricas e fotométricas
P1f	Pequeno	Etapa única	Geométricas e fotométricas

Tabela 4: Resumo dos modelos treinados com diferentes processos e tipos de aumentação.

Para avaliar o desempenho do classificador optamos por utilizar a sensibilidade condicionada a uma especificidade mínima de 95%, a mesma métrica usada para avaliar a classificação binária na competição JustRAIGS. Essa escolha pode ser justificada pelo contexto clínico da aplicação: na triagem do glaucoma, queremos minimizar o número de falsos positivos, evitando sobrecarregar o sistema de saúde com encaminhamentos desnecessários. A alta especificidade garante que os pacientes encaminhados realmente apresentem sinais relevantes de glaucoma. Ao mesmo tempo, queremos maximizar a sensibilidade, isto é, a capacidade de detectar os casos verdadeiramente positivos e garantir que não estamos deixando de encaminhá-los.

No contexto da identificação do glaucoma, a precisão é fortemente impactada pelo grande desbalanceamento entre as classes. Por mais que tenhamos um percentual pequeno de falsos positivos em relação ao número total de casos, em números absolutos ainda vamos ter muitos casos em comparação à quantidade de verdadeiros positivos, resultando em uma precisão baixa. Portanto, apesar de reportá-la, optamos por não nos concentrar nessa métrica.

A Tabela 5 apresenta os resultados obtidos pelos modelos no conjunto de validação. O modelo G2, de maior porte, teve o pior desempenho, indicando que o aumento do tamanho da rede, isoladamente, não trouxe melhorias. A adição de aumentações fotométricas (modelo M2f) resultou em pequenos ganhos. Já a adoção de treinamento em etapa única (modelos M1f e P1f) proporcionou melhorias mais expressivas, com destaque para o modelo M1f, que alcançou a melhor sensibilidade (93,87%). A redução do tamanho da rede no modelo P1f implicou em leve queda nas métricas, mas ainda superando os primeiros modelos avaliados.

Nome	Sens. em 95%	AUC-ROC	Especif.	Sensitiv.	Precisão
M2	0,9023	0,9764	0,9665	0,8448	0,8681
G2	0,8851	0,9724	0,9675	0,8084	0,8665
M2f	0,9061	0,9746	0,9675	**0,8506**	0,8723
M1f	**0,9387**	**0,9807**	**0,9780**	**0,8506**	**0,9098**
P1f	0,9215	0,9806	0,9745	0,8391	0,8957

Tabela 5: Resultados dos modelos avaliados no conjunto de validação.

4.5 Classificação multi-rótulo

Após a classificação binária entre casos referenciáveis (RG) e não referenciáveis (NRG), foi desenvolvido um segundo classificador com o objetivo de identificar as razões que justificam a marcação de uma imagem como RG, seguindo as anotações adicionais propostas pelo conjunto de dados JustRAIGS.

Este classificador foi treinado apenas com as imagens rotuladas como RG no conjunto de treinamento. Uma mesma imagem pode estar associada à mais de uma razão, caracterizando o problema como uma tarefa de classificação multi-rótulo (*multi-label classification*) com dez saídas, correspondentes aos dez possíveis rótulos adicionais anotados pelos avaliadores.

A arquitetura do classificador multi-rótulo foi construída a partir da mesma base do classificador binário, utilizando uma ResNet50 pré-treinada no ImageNet com camadas superiores substituídas por uma *GlobalAveragePooling2D* e camadas totalmente conectadas de tamanhos 1024, 256 e 10, conforme descrito na Seção 4.4. A diferença para o classificador binário está na camada de saída que possui dez unidades, cada uma com função de ativação sigmoide, permitindo a previsão independente de cada razão.

O treinamento foi realizado em uma única etapa, com taxa de aprendizagem de 10^{-5} e otimizador ADAM, empregando as mesmas técnicas de aumentação de dados e estratégias de regularização aplicadas na tarefa binária, incluindo *dropout* e parada antecipada (*early stopping*).

Como descrito na Seção 4.1, no conjunto de dados JustRAIGS, é possível que os avaliadores, mesmo concordando em referenciar uma imagem para glaucoma, discordem

nas anotações adicionais que justificam essa decisão. O JustRAIGS não implementou um mecanismo de resolução de conflitos entre avaliadores para essas anotações, disponibilizando apenas as respostas individuais. Dessa forma, propusemos uma estratégia de construção de anotações finais contínuas, conhecidas como *soft-labels*, que refletem a incerteza inerente às anotações discordantes.

O processo para definição do valor final de cada anotação adicional foi estabelecido da seguinte maneira:

- Consideram-se A1 e A2 como os dois avaliadores iniciais, e A3 como o especialista em glaucoma acionado em caso de discordância na anotação principal.
- **Se A1 e A2 concordam na anotação principal:**
  Para cada anotação adicional:
  - Se ambos marcaram 0, o valor final atribuído é 0.
  - Se ambos marcaram 1, o valor final atribuído é 1.
  - Se discordam, o valor final atribuído é 0,6, refletindo a incerteza entre os avaliadores.
- **Se houve discordância na anotação principal e foi necessária a avaliação de A3:**
  - **Caso A3 concorde em referenciar a imagem:**
    Para cada anotação adicional:
    - Se o avaliador inicial (A1 ou A2) e A3 concordam em 0, o valor final é 0.
    - Se concordam em 1, o valor final é 1.
    - Se apenas o avaliador inicial marcou 1, o valor final é 0,25.
    - Se apenas A3 marcou 1, o valor final é 0,9, atribuindo maior peso à opinião do especialista.
  - **Caso A3 opte por não referenciar a imagem:**
    - Para cada anotação adicional em que um dos avaliadores iniciais marcou 1, é atribuído o valor final de 0,05, representando um indício fraco da ocorrência da característica.
- **Tratamento de casos de imagens marcadas como não avaliáveis (*ungradable*):**
  Considerando que A1 marcou a imagem como *ungradable*

- **Se A2 e A3 concordam em RG:**
  - Usamos a regra já descrita anteriormente
- **Se A2 e A3 discordam**
  - Se A3 marcou RG: utilizamos as anotações de A3 diretamente
  - Se A3 marcou NRG: as anotações de A2 recebem o valor de 0,05

Essa estratégia visa refletir a incerteza das anotações entre os avaliadores e o peso maior da opinião do especialista em casos de discordância.

Para a avaliação do desempenho do classificador multi-rótulo, utilizamos uma distância de Hamming modificada. A métrica tradicional considera a quantidade de posições que são diferentes entre dois vetores. No entanto, devido a existência de discordâncias entres os avaliadores nas anotações adicionais, optamos por ignorar no cálculo os rótulos para os quais houve desacordo entre os avaliadores. Dessa forma, a distância foi calculada somente para os rótulos em que houve consenso e normalizada pela quantidade deles.

No conjunto de validação, o modelo atingiu uma distância de Hamming média de 0,12247. Para fins de comparação, avaliamos um classificador trivial que prevê a ausência de todas as razões (todos os rótulos iguais a zero) para todas as imagens. Esse classificador *baseline* obteve uma distância de Hamming média de 0,28819, demonstrando que o modelo treinado superou a estratégia básica.

4.6 Recursos computacionais

Todos os modelos foram implementados em Python com Keras utilizando o TensorFlow v2.18.0 de backend, com exceção do YOLO (Ultralytics) que utiliza o PyTorch. As versões dos pacotes eram: Python 3.10.12, Keras 3.6.0, TensorFlow 2.18.0, Ultralytics 8.3.18. Os códigos foram executados com o sistema operacional Pop!_OS 22.04, kernel Linux 6.9.3, contendo instalado o driver da NVIDIA 560.35.03 e as bibliotecas NVIDIA CUDA 12.6 e NVIDIA cuDNN 9.5.1. O computador estava equipado com um processador Intel i7-10700K, 32 GB de RAM e uma GPU NVIDIA GeForce RTX 3070.

5 Resultados

Esta seção apresenta os resultados obtidos pelo sistema proposto no conjunto de teste definido na Seção 4.1.3, que não foi utilizado em nenhuma etapa de desenvolvimento. Este

conjunto visa avaliar a capacidade de generalização dos modelos, isto é, seu desempenho em casos nunca antes vistos.

Inicialmente, o segmentador foi aplicado a todas as 20.285 imagens do conjunto de teste. Ele retornou exatamente uma detecção de disco óptico em 20.247 imagens (99,81%), mais de uma detecção em 19 imagens (0,09%) e nenhuma detecção em outras 19 imagens (0,09%). A pontuação média das detecções foi de 0,922, com mediana de 0,927. Entre as imagens anotadas como glaucoma, apenas duas não tiveram detecção.

As imagens com exatamente uma detecção foram então avaliadas pelo classificador binário. Diferentes modelos, treinados em diversas configurações, foram testados. O modelo P1f, treinado em etapa única com todas as técnicas de aumentação e utilizando a menor arquitetura entre as avaliadas, apresentou o melhor desempenho no conjunto de teste. Este modelo havia obtido o segundo melhor desempenho na etapa de validação. No teste, atingiu sensitividade de 0,8972 em 95% de especificidade mínima, AUC-ROC de 0,9781, especificidade de 0,9691, sensitividade de 0,8543 e precisão de 0,4789. A Tabela 6 apresenta o resumo das métricas para este e outros modelos avaliados.

Tabela 6: Resultados dos modelos avaliados no conjunto de teste.

Nome	**Sens. em 95%**	**AUC-ROC**	**Especif.**	**Sensitiv.**	**Precisão**
M2	0,8773	0,9726	0,9643	0,8451	0,4408
M2f	0,8758	0,9741	0,9611	0,8420	0,4188
M1f	0,8819	0,9765	0,9667	0,8390	0,4558
P1f	**0,8972**	**0,9781**	**0,9691**	**0,8543**	**0,4789**

É importante observar que a precisão obtida no conjunto de teste (0,4789) é substancialmente menor do que aquela observada na validação (0,8957), o que pode causar estranheza à primeira vista. No entanto, essa diferença é explicada pela proporção entre as classes nos conjuntos: enquanto o conjunto de validação foi balanceado para conter aproximadamente 1 imagem RG a cada 4 NRG, no conjunto de teste a proporção realista é de 1 RG para 31 NRG, a mesma do JustRAIGS como um todo. Como consequência, mesmo uma pequena taxa de falsos positivos resulta em um grande número absoluto de erros em comparação à quantidade de verdadeiros positivos, o que reduz a precisão.

Apesar disso, a sensitividade em 95% de especificidade foi mantida em um patamar elevado (89,72%), o que reforça a robustez do modelo na triagem de glaucoma mesmo sob forte desbalanceamento.

Em seguida, dentre as 20.247 imagens com detecção única, 652 possuíam anotação positiva para glaucoma. Estas foram avaliadas pelo classificador multi-rótulo, que obteve distância de Hamming de 0,12960, valor ligeiramente superior ao obtido na validação (0,12247), mas consideravelmente inferior ao *baseline* (0,28819) obtido por um classificador ingênuo.

Os resultados obtidos evidenciam a capacidade do sistema proposto em realizar a triagem de glaucoma e justificar suas decisões de forma interpretável. A Tabela 7 apresenta uma comparação entre os resultados deste trabalho e aqueles reportados por outros estudos com o conjunto de dados JustRAIGS, conforme discutido na Seção 3.6.1.

Autor(es)	**Arquitetura(s)**	**Estratégia adicional**	**Sensitiv. 95% Esp.**	**Dist. de Hamming**
Galdran e Ballester [7]	ResNet50	Suavização de rótulos	94,33%*	0,1440*
Zhang et al. [38]	Ensemble ResNet50 + Transformers	-	87,00%	0,239
Kubrak [13]	YOLOv8 + Vision Transformer	Segmentação do disco	90,90%	0,1280
Lin et al. [17]	YOLOv8 + Vision Transformer	Segmentação do disco	85,70%	0,1250
Presente trabalho	YOLOv8 + ResNet50	Segmentação do disco + Suavização de rótulos	89,72%**	0,1296**

Nota: os valores foram obtidos em conjuntos de teste distintos e, portanto, não são diretamente comparáveis.

(*) Métricas obtidas em subconjuntos de validação.

(**) Métricas obtidas em subconjunto de teste independente.

Demais valores foram reportados no conjunto de teste da competição na etapa de validação.

Tabela 7: Comparativo do presente trabalho com outros que utilizaram o JustRAIGS.

6 Conclusões e Trabalhos Futuros

Este trabalho propôs o desenvolvimento de um sistema baseado em aprendizagem profunda para a detecção automatizada de casos de glaucoma em imagens de fundo de olho, utilizando a segmentação do disco óptico, a classificação binária e a identificação multi-rótulo das características clínicas justificadoras. Foram empregados redes neurais convolucionais modernas, técnicas de aumento de dados e estratégias de regularização no treinamento dos modelos.

No conjunto de teste definido, o sistema atingiu uma sensitividade de 89,72% em 95% de especificidade mínima e uma AUC-ROC de 0,9781 na tarefa de classificação binária. Para a tarefa de classificação multi-rótulo, obteve distância de Hamming de 0,12960, superando significativamente o *baseline* de 0,28819 correspondente a um classificador ingênuo.

Comparando com trabalhos relacionados desenvolvidos com o mesmo conjunto de dados JustRAIGS, o desempenho do sistema posiciona-se de maneira competitiva. Na tarefa de classificação binária, Kubrak [13] atingiu sensitividade de 90,90% em 95% de especificidade no conjunto de testes da competição, enquanto Lin et al. [17] obtiveram 85,70%. Nosso sistema obteve 89,72% no conjunto de teste interno. Na tarefa multi-rótulo, Kubrak reportou distância de Hamming de 0,1280, Lin et al. [17] reportaram 0,1250, enquanto Zhang et al. [38] reportaram 0,239. O sistema proposto apresentou desempenho próximo desses trabalhos, com distância de Hamming de 0,12960, utilizando uma arquitetura mais simples e sem a aplicação de *ensembles* ou *transformers*.

Cabe ressaltar que os resultados apresentados não são diretamente comparáveis, uma vez que foram avaliados em conjuntos de teste distintos. Apesar disso, o uso de um conjunto de teste separado e não utilizado durante o treinamento permite uma comparação aproximada, fornecendo uma referência válida sobre o posicionamento do sistema proposto frente aos métodos existentes.

Desta maneira, evidenciamos que os objetivos propostos foram atingidos:

- Desenvolvemos um método de segmentação do disco óptico com taxa de sucesso de 99,81%;
- Apresentamos um classificador binário que atingiu desempenho robusto para a triagem de glaucoma;
- Apresentamos um classificador multi-rótulo que atingiu desempenho comparável

com os disponíveis na literatura no apontamento das razões clínicas que justificam a decisão.

Entre as limitações do trabalho, destacamos a avaliação do classificador multi-rótulo. Foi utilizada a distância de Hamming com o objetivo de tornar a avaliação compatível com a proposta da competição, porém não está claro se esta é a métrica mais adequada para o contexto clínico. Possivelmente, outras métricas poderiam refletir melhor o desempenho em cenários de triagem. Destacamos também que o sistema foi avaliado apenas em um único conjunto de dados, sem validação externa em outras bases ou contextos clínicos.

Como trabalhos futuros, propõe-se:

- Explorar métodos de interpretabilidade visual para explicar as decisões do modelo;
- Realizar testes em bases de dados adicionais e avaliações clínicas no mundo real.

Os resultados obtidos indicam que sistemas baseados em aprendizagem profunda apresentam grande potencial como ferramenta de apoio à triagem de glaucoma, contribuindo para o diagnóstico precoce e a redução da cegueira potencialmente evitável.

Referências

[1] Ahmed Almazroa. Retinal fundus images for glaucoma analysis: the RIGA dataset, 2018.

[2] Christopher M. Bishop. *Pattern Recognition and Machine Learning*. Springer, 2006.

[3] X. Chen, Y. Xu, D. W. Kee Wong, T. Y. Wong, and J. Liu. Glaucoma detection based on deep convolutional neural network. In *2015 37th Annual International Conference of the IEEE Engineering in Medicine and Biology Society (EMBC)*, pages 715–718, Aug 2015.

[4] Andres Diaz-Pinto, Sandra Morales, Valery Naranjo, Thomas Köhler, Jose M. Mossi, and Amparo Navea. CNNs for automatic glaucoma assessment using fundus images: an extensive validation. *BioMedical Engineering OnLine*, 18(1):29, December 2019.

[5] Marcos Vinícius Dos Santos Ferreira, Antonio Oseas de Carvalho Filho, Alcilene Dalília De Sousa, Aristófanes Corrêa Silva, and Marcelo Gattass. Convolutional neural network and texture descriptor-based automatic detection and diagnosis of glaucoma. *Expert Systems with Applications*, 110:250–263, November 2018.

[6] Francisco José Fumero Batista, Tinguaro Diaz-Aleman, Jose Sigut, Silvia Alayon, Rafael Arnay, and Denisse Angel-Pereira. Rim-one dl: A unified retinal image database for assessing glaucoma using deep learning. *Image Analysis and Stereology*, 39(3):161–167, Nov. 2020.

[7] Adrian Galdran and Miguel A. González Ballester. Data-centric label smoothing for explainable glaucoma screening from eye fundus images, 2024.

[8] Ian Goodfellow, Yoshua Bengio, and Aaron Courville. *Deep Learning*. MIT Press, 2016.

[9] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition, 2015.

[10] Andrew G. Howard, Menglong Zhu, Bo Chen, Dmitry Kalenichenko, Weijun Wang, Tobias Weyand, Marco Andreetto, and Hartwig Adam. Mobilenets: Efficient convolutional neural networks for mobile vision applications, 2017.

[11] Glenn Jocher, Jing Qiu, and Ayush Chaurasia. Ultralytics YOLO, January 2023.

[12] Alex Krizhevsky, Ilya Sutskever, and Geoffrey E Hinton. Imagenet classification with deep convolutional neural networks. In F. Pereira, C.J. Burges, L. Bottou, and K.Q. Weinberger, editors, *Advances in Neural Information Processing Systems*, volume 25. Curran Associates, Inc., 2012.

[13] Tomasz Kubrak. Automated detection of glaucoma and diagnostic features for justraigs challenge. In *2024 IEEE International Symposium on Biomedical Imaging (ISBI)*, pages 1–3, 2024.

[14] Yann LeCun, Léon Bottou, Yoshua Bengio, and Patrick Haffner. Gradient-based learning applied to document recognition. *Proceedings of the IEEE*, 86(11):2278–2324, 1998.

[15] Hans G. Lemij, Coen de Vente, Clara I. Sánchez, and Koen A. Vermeer. Characteristics of a Large, Labeled Data Set for the Training of Artificial Intelligence for Glaucoma Screening with Fundus Photographs. *Ophthalmology Science*, 3(3):100300, September 2023.

[16] Tao Li, Wang Bo, Chunyu Hu, Hong Kang, Hanruo Liu, Kai Wang, and Huazhu Fu. Applications of deep learning in fundus images: A review. *Medical Image Analysis*, 69:101971, April 2021.

[17] Hui Lin, Charilaos Apostolidis, and Aggelos K. Katsaggelos. Brighteye: Glaucoma screening with color fundus photographs based on vision transformer. In *2024 IEEE International Symposium on Biomedical Imaging (ISBI)*, pages 1–4, 2024.

[18] Geert Litjens, Thijs Kooi, Babak Ehteshami Bejnordi, Arnaud Arindra Adiyoso Setio, Francesco Ciompi, Mohsen Ghafoorian, Jeroen A.W.M. van der Laak, Bram van Ginneken, and Clara I. Sánchez. A survey on deep learning in medical image analysis. *Medical Image Analysis*, 42:60–88, 2017.

[19] Hanruo Liu, Liu Li, I. Michael Wormstone, et al. Development and Validation of a Deep Learning System to Detect Glaucomatous Optic Neuropathy Using Fundus Photographs. *JAMA Ophthalmology*, 137(12):1353–1360, December 2019.

[20] Christos Matsoukas, Johan Fredin Haslum, Moein Sorkhei, Magnus Söderberg, and Kevin Smith. What makes transfer learning work for medical images: Feature reuse & other factors. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 9225–9234, June 2022.

[21] Tom M. Mitchell. *Machine Learning.* McGraw-Hill, 1997.

[22] Rotterdam Ophthalmic Institute; Rotterdam Eye Hospital; Rotterdam; Netherlands. Justraigs challenge training data set, January 2024.

[23] Viet Anh Nguyen. AnyLabeling - Effortless data labeling with AI support.

[24] Kevin P. Noronha, U. Rajendra Acharya, K. Prabhakar Nayak, Roshan Joy Martis, and Sulatha V. Bhandary. Automated classification of glaucoma stages using higher order cumulant features. *Biomedical Signal Processing and Control*, 10:174–183, 2014.

[25] José Ignacio Orlando, Huazhu Fu, João Barbosa Breda, et al. Refuge challenge: A unified framework for evaluating automated methods for glaucoma assessment from fundus photographs. *Medical Image Analysis*, 59:101570, January 2020.

[26] Olga Russakovsky, Jia Deng, Hao Su, Jonathan Krause, Sanjeev Satheesh, Sean Ma, Zhiheng Huang, Andrej Karpathy, Aditya Khosla, Michael Bernstein, Alexander C. Berg, and Li Fei-Fei. ImageNet Large Scale Visual Recognition Challenge. *International Journal of Computer Vision (IJCV)*, 115(3):211–252, 2015.

[27] Stuart J. Russell and Peter Norvig. *Artificial Intelligence: A Modern Approach.* Prentice Hall, 3rd edition, 2010.

[28] Karen Simonyan and Andrew Zisserman. Very deep convolutional networks for large-scale image recognition. In *International Conference on Learning Representations*, 2015.

[29] Jayanthi Sivaswamy, Subbaiah Krishnadas, Arunava Chakravarty, Gopal Joshi, and Ujjwal. A comprehensive retinal image dataset for the assessment of glaucoma from the optic nerve head analysis. *JSM Biomed Imaging Data Pap*, 2, 01 2015.

[30] Jayanthi Sivaswamy, Subbaiah Krishnadas, Gopal Joshi, Madhulika Jain, and A. Tabish. Drishti-gs: Retinal image dataset for optic nerve head(onh) segmentation. In

2014 IEEE 11th International Symposium on Biomedical Imaging, ISBI 2014, pages 53–56, 04 2014.

[31] Jaimie D Steinmetz, Rupert R A Bourne, Paul Svitil Briant, Seth R Flaxman, Hugh R B Taylor, Jost B Jonas, et al. Causes of blindness and vision impairment in 2020 and trends over 30 years, and prevalence of avoidable blindness in relation to VISION 2020: the Right to Sight: an analysis for the Global Burden of Disease Study. *The Lancet Global Health*, 9(2):e144–e160, February 2021.

[32] Christian Szegedy, Wei Liu, Yangqing Jia, Pierre Sermanet, Scott Reed, Dragomir Anguelov, Dumitru Erhan, Vincent Vanhoucke, and Andrew Rabinovich. Going deeper with convolutions. In *Proceedings of the IEEE conference on computer vision and pattern recognition (CVPR)*, pages 1–9, 2015.

[33] Nicholas Y.Q. Tan, David S. Friedman, Ingeborg Stalmans, Iqbal Ike K. Ahmed, and Chelvin C.A. Sng. Glaucoma screening: where are we and where do we need to go? *Current Opinion in Ophthalmology*, 31(2), 2020.

[34] Yih-Chung Tham, Xiang Li, Tien Y. Wong, Harry A. Quigley, Tin Aung, and Ching-Yu Cheng. Global Prevalence of Glaucoma and Projections of Glaucoma Burden through 2040: A Systematic Review and Meta-Analysis. *Ophthalmology*, 121(11):2081–2090, November 2014. Publisher: Elsevier.

[35] Robert N Weinreb and Peng Tee Khaw. Primary open-angle glaucoma. *The Lancet*, 363(9422):1711–1720, May 2004.

[36] Robert N. Weinreb, Christopher K. S. Leung, Jonathan G. Crowston, Felipe A. Medeiros, David S. Friedman, Janey L. Wiggs, and Keith R. Martin. Primary open-angle glaucoma. *Nature Reviews Disease Primers*, 2(1):16067, September 2016.

[37] World Health Organization. *World report on vision*. World Health Organization, Geneva, 2019.

[38] Philippe Zhang, Yihao Li, Jing Zhang, Weili Jiang, Pierre-Henri Conze, Mathieu Lamard, Gwenolé Quellec, and Mostafa El Habib Daho. Detection and classification of glaucoma in the justraigs challenge: Achievements in binary and multilabel clas-

sification. In *2024 IEEE International Symposium on Biomedical Imaging (ISBI)*, pages 1–4, 2024.

[39] Zhuo Zhang, Feng Shou Yin, Jiang Liu, Wing Kee Wong, Ngan Meng Tan, Beng Hai Lee, Jun Cheng, and Tien Yin Wong. ORIGA(-light): an online retinal fundus image database for glaucoma analysis and research. *Annual International Conference of the IEEE Engineering in Medicine and Biology Society. IEEE Engineering in Medicine and Biology Society. Annual International Conference*, 2010:3065–3068, 2010.