UNIVERSIDADE FEDERAL DO MARANHÃO

UNIVERSIDADE FEDERAL DO PIAUÍ

Doutorado em Ciência da Computação Associação UFMA/UFPI

Marcos Melo Ferreira

Classificação do estágio de glaucoma usando dados multimodais

Orientador: Prof. Dr. Geraldo Braz Junior

Co-orientador: Prof. Dr. António Cunha

São Luís - MA
Setembro, 2025

Marcos Melo Ferreira

Classificação do estágio de glaucoma usando dados multimodais

TESE DE DOUTORADO

Tese apresentada como requisito para obtenção do título de Doutor em Ciência da Computação, ao Doutorado em Ciência da Computação, Associação UFMA/UFPI.

Orientador: Prof. Dr. Geraldo Braz Junior
Co-orientador: Prof. Dr. António Cunha

São Luís - MA
Setembro, 2025

Ficha gerada por meio do SIGAA/Biblioteca com dados fornecidos pelo(a) autor(a).
Diretoria Integrada de Bibliotecas/UFMA

Ferreira, Marcos.
Classificação do estágio de glaucoma usando dados multimodais / Marcos Ferreira. - 2025.
95 f.

Coorientador(a) 1: António Cunha.
Orientador(a): Geraldo Braz.
Tese (Doutorado) - Programa de Pós-graduação Doutorado em Ciência da Computação - Associação UFMA/UFPI, Universidade Federal do Maranhão, Sao Luis, 2025.

1. Classificação de Estágios de Glaucoma. 2. Retinografia. 3. Tomografia de Coerência Óptica. 4. Deep Learning. 5. Modelos Multimoais. I. Braz, Geraldo. II. Cunha, António. III. Título.

Documento assinado digitalmente
MARCOS MELO FERREIRA
Data: 16/10/2025 16:27:39-0300
Verifique em https://validar.iti.gov.br

Marcos Melo Ferreira

Classificação do estágio de glaucoma usando dados multimodais

A presente Tese de Doutorado foi avaliada e aprovada por banca examinadora composta pelos seguintes membros:

Prof. Dr. Geraldo Braz Junior
Orientador
Universidade Federal do Maranhão

Prof. Dr. António Cunha
Co-orientador
Universidade Federal do Maranhão

Prof. Dr. Francesco Renna
Examinador Externo
Universidade de Porto

Prof. Dr. Paulo Ivson Netto Santos
Examinador Externo
Pontifícia Universidade Católica do Rio de Janeiro

Rodrigo de Melo Souza Veras
Examinador Interno
Universidade Federal do Piauí

Prof. Dr. João Dallyson Sousa de Almeida
Examinador Interno
Universidade Federal do Maranhão

Certificamos que esta é a versão original e final da Tese de Doutorado que foi julgada adequada para obtenção do título de Doutor em Ciência da Computação.

Documento assinado digitalmente
GERALDO BRAZ JUNIOR
Data: 20/10/2025 11:26:20-0300
Verifique em https://validar.iti.gov.br

Prof. Dr. Geraldo Braz Junior
Orientador

Prof. Dr. Anselmo Cardoso de Paiva
Coordenador

São Luís - MA, 19 de Setembro de 2025

Aos meus pais, irmãos, professores e amigos que sempre me apoiaram e deram todo o suporte necessário para a conclusão desta etapa.

Agradecimentos

Esta proposta de tese de doutorado é resultado de muito trabalho e dedicação e é importante demonstrar os meus sinceros agradecimentos aos que me ajudaram nesta etapa da minha vida.

Primeiramente, a Deus, pela saúde e determinação para superar todas as dificuldades que se apresentaram.

Aos meus pais, pois sempre acreditaram na minha capacidade, sempre me incentivaram e me apoiaram, estando ao meu lado, me dando amor e suporte durante toda a minha vida.

A todos os meus irmãos, sou muito grato pela amizade e confiança.

Aos professores do programa de pós-graduação, em especial ao meu orientador, professor Geraldo, e ao meu coorientador, professor António, pela oportunidade, por estarem sempre disponíveis para orientar e ajudar, pela paciência, pelo incentivo e por tudo o que me ensinaram.

Aos meus companheiros de pesquisa do Viplab, da Universidade Federal do Maranhão, sempre prontos a ajudar e sugerir soluções para os problemas que se apresentaram.

Aos meus colegas de investigação do Laboratório de pesquisa e apoio ao ensino, da Universidade de Trás-os-Montes e Alto Douro (UTAD), pelo acolhimento, pela amizade, pelas sugestões e todo o suporte oferecido durante minha estadia em Portugal.

Ao Instituto Federal de Educação, Ciência e Tecnologia do Maranhão - Campus São José de Ribamar, muito obrigado por acreditar no meu potencial e me liberar para seguir este importante caminho acadêmico, que contribui grandemente para o meu desenvolvimento profissional e pessoal.

À Coordenação de Aperfeiçoamento de Pessoal de Nível Superior - Brasil (CAPES) - por possibilitar uma oportunidade de desenvolver a pesquisa científica no exterior.

Agradeço a todos que de alguma forma contribuíram para minha formação acadêmica.

“A persistência é amiga da conquista. Se você quer chegar aonde a maioria não chega, faça o que a maioria não faz.”

Bill Gates.

Resumo

O glaucoma é a principal causa de cegueira irreversível no mundo. Seu diagnóstico precoce é desafiador devido à ausência de sintomas nos estágios iniciais, à necessidade de análise de múltiplos exames por profissionais especializados e ao baixo conhecimento da população sobre a doença. Embora a perda visual causada pelo glaucoma seja irreversível, sua progressão pode ser retardada quando identificada precocemente. Nesse contexto, métodos baseados em deep learning têm se mostrado promissores em tarefas de processamento de imagens médicas, como classificação e segmentação, oferecendo suporte potencial ao diagnóstico clínico. Neste trabalho, foi desenvolvido um método para classificação dos estágios do glaucoma a partir da utilização de retinografias e volumes de OCT. O método emprega uma arquitetura multimodal baseada em redes convolucionais e explora diferentes estratégias de fusão, tanto de mapas de características quanto de predições, com o objetivo de integrar de forma eficaz as modalidades. Além disso, foram investigadas regiões específicas de interesse — o nervo óptico, nas retinografias, e as camadas retinianas, nos volumes de OCT — como forma de aprimorar a representação dos dados e melhorar a acurácia da classificação. Os experimentos realizados demonstraram que os modelos multimodais alcançaram desempenhos superiores em relação aos unimodais, atingindo como melhor resultado um valor de Kappa de 0,88, o que indica um alto nível de concordância do método proposto em relação às avaliações de especialistas. Adicionalmente, os resultados evidenciaram que a retinografia exerce maior influência do que os volumes de OCT no processo de classificação, enquanto a captura direcionada das camadas da retina mostrou-se uma abordagem promissora para aumentar a precisão do modelo. De forma geral, o método proposto demonstrou potencial significativo como ferramenta de apoio à decisão clínica, contribuindo para o avanço de sistemas automatizados de diagnóstico e para a detecção precoce e precisa do glaucoma.

Palavras-chave: Classificação de estágios de glaucoma, Retinografia, Tomografia de coerência óptica, Deep learning, Redes neurais convolucionais, Modelos multimodais.

Abstract

Glaucoma is the leading cause of irreversible blindness worldwide. Its early diagnosis is challenging due to the absence of symptoms in the initial stages, the need for multiple exams to be analysed by specialised professionals, and the general lack of awareness about the disease among the population. Although the visual loss caused by glaucoma is irreversible, its progression can be slowed if the disease is detected in its early stages. In this context, deep learning methods have demonstrated promising results in medical image processing tasks, including classification and segmentation, offering potential support for clinical diagnosis. In this work, we developed a method for glaucoma stage classification that combines fundus photographs and OCT volumes. The method employs a multimodal convolutional architecture and explores various fusion strategies, both at the feature map and prediction levels, aiming to integrate multimodal information effectively. Additionally, specific regions of interest were investigated — the optic nerve in fundus photographs and the retinal layers in OCT volumes — to improve data representation and enhance classification accuracy. The experiments demonstrated that multimodal models outperformed unimodal approaches, achieving a Kappa score of 0.88, which indicates a high level of agreement of the proposed method with specialist assessments. Moreover, the results showed that fundus photography has a greater influence than OCT volumes in the classification process. At the same time, the targeted capture of retinal layers proved to be a promising strategy for further improving accuracy. Overall, the proposed method demonstrated significant potential as a clinical decision support tool, contributing to the advancement of automated diagnostic systems and enabling earlier and more accurate glaucoma detection.

Keywords: Glaucoma grading stages, Retinography, Optical coherence tomography, Deep learning, Convolutional neural networks, Multimodal models.

Lista de ilustrações

Lista de tabelas

Lista de abreviaturas e siglas

AF	Angiografia Fluoresceínica
AUC	*Area Under The Curve*
CBAM	*Convolutional Block Attention Module*
CBO	Conselho Brasileiro de Oftalmologia
CFP	*Color Fundus Photography*
CLAHE	*Contrast Limited Adaptive Histogram Equalization*
CNN	*Convolutional Neural Network*
DenseNet	*Dense Convolutional Network*
DMRI	Degeneração Macular Relacionada à Idade
EMD	Edema Macular Diabético
ETDRS	*Early Treatment Diabetic Retinopathy Study*
FDA	*Food And Drug Administration*
GAMMA	*Glaucoma Grading From Multi-Modality Images*
GCIPL	*Ganglion Cell-Inner Plexiform Layer*
GRAD-CAM	*Gradient-Weighted Class Activation Mapping*
HM	Hemorragias
IBGE	Instituto Brasileiro de Geografia e Estatística
IBOPE	Instituto Brasileiro de Opinião Pública e Estatística
IDF	*International Diabetes Federation*
ILSVRC	*Imagenet Large-Scale Visual Recognition Challenge*
LR	*Learning Rate*
MA	Microaneurismas
MIFS	*Multi Instance Feature Selection*

NVE	Neovascularização
OCT	*Optical Coherence Tomography*
OCTA	*Optical Coherence Tomography Angiography*
OMS	Organização Mundial Da Saúde
PIO	Pressão Intraocular
RDNP	Retinopatia Diabética Não Proliferativa
RDP	Retinopatia Diabética Proliferativa
ReLU	*Rectified Linear Units*
ResNet	*Residual Networks*
RNFL	*Retinal Nerve Fiber Layer*
SE	*Squeeze-and-Excitation*
SGD	*Stochastic Gradient Descent*
SMBO	*Sequential Model-Based Optimization*
SOTA	*State Of The Art*
SVM	*Support Vector Machine*
TC	Totalmente Conectadas
UWF	*Ultra-Widefield*
VGG	*Visual Geometry Group*
WGBF	*Weighted Gaussian Blur Fundus*

Sumário

1 Introdução

O glaucoma é uma das principais causas de deficiência visual e de cegueira irreversível em todo o mundo (WHO, 2019). Trata-se de uma doença ocular crônica e progressiva, caracterizada por danos graduais ao nervo óptico, frequentemente associados ao aumento da pressão intraocular PIO. Esses danos levam à perda bilateral e irreversível da visão, que pode evoluir até a cegueira total se não houver diagnóstico e intervenção precoces. Além da pressão elevada, fatores como idade avançada, histórico familiar e etnia estão relacionados ao risco de desenvolvimento da doença (SARHAN; ROKNE; ALHAJJ, 2019).

Segundo o Conselho Brasileiro de Oftalmologia CBO, estimativas do Instituto Brasileiro de Geografia e Estatística (IBGE) indicam que aproximadamente 1,6 milhão de pessoas são cegas no país (CBO, 2020), sendo o glaucoma uma das principais causas. A doença apresenta um desafio significativo para a saúde pública devido ao seu caráter silencioso: cerca de 80% dos portadores não apresentam sintomas na fase inicial da doença (CBO, 2017). A pesquisa “Um novo olhar para o glaucoma no Brasil”, realizada em junho de 2020 pelo IBOPE Inteligência, revelou que 41% dos entrevistados não sabiam o que é glaucoma e 53% desconheciam que ele possui a maior probabilidade de levar a um quadro de cegueira irreversível. Em escala global, em 2020, cerca de 3,61 milhões de indivíduos eram cegos e quase 4,14 milhões tinham deficiência visual em decorrência do glaucoma (STUDY et al., 2024). Além disso, estima-se que aproximadamente 50% dos indivíduos com glaucoma desconheçam sua condição, sendo esse percentual possivelmente ainda maior em países subdesenvolvidos (SOH et al., 2021). Segundo estimativas da Organização Mundial da Saúde (OMS), publicadas em 2017, o número de pessoas com glaucoma poderá alcançar cerca de 122 milhões até 2040 (WHO, 2019). Embora a perda visual causada pela doença seja irreversível, sua progressão pode ser significativamente retardada por meio de diagnóstico e tratamento oportunos. Ainda de acordo com a OMS (WHO, 2019), entre os indivíduos diagnosticados e tratados precocemente, 11% relataram deficiência visual moderada ou grave, ou cegueira resultante de formas mais avançadas do glaucoma.

Para prevenir danos irreversíveis, é essencial que o glaucoma seja diagnosticado em estágios iniciais, possibilitando intervenções que impeçam sua evolução. Contudo, por ser assintomática em suas fases iniciais, a doença é frequentemente detectada em estágios mais avançados, quando já há danos estruturais relevantes (SARHAN; ROKNE; ALHAJJ, 2019). O rastreamento e o diagnóstico podem ser realizados por meio de exames de imagem, como a retinografia - fotografias coloridas de fundo de olho (CFP, do inglês *Color Fundus Photography*) - e tomografia de coerência óptica (OCT, do inglês

Optical Coherence Tomography), além de outros exames complementares, como a tomografia por coerência óptica com angiografia (OCTA) e a angiografia fluoresceínica (AN et al., 2019).

A retinografia é um exame padrão para a avaliação do disco óptico, escavação e vasos sanguíneos, sendo rápida, não invasiva e de ampla aplicação clínica. Por outro lado, a OCT fornece cortes seccionais da retina em alta resolução, permitindo a análise detalhada de suas camadas, como a camada de fibras nervosas da retina (RNFL, do inglês *Retinal Nerve Fiber Layer*) e a camada de células ganglionares (GCIPL, do inglês *Ganglion celular Inner plexiform layer*), frequentemente afetadas nos estágios iniciais do glaucoma. Assim, o uso conjunto dessas modalidades amplia a capacidade diagnóstica, fornecendo informações complementares que auxiliam na avaliação da estrutura retiniana e na detecção de alterações morfológicas sutis.

Entretanto, o rastreamento populacional em larga escala é limitado pela escassez de especialistas: no Brasil, por exemplo, a média é de 1 oftalmologista para cada 10.875 habitantes, proporção que chega a 1:19.512 na região Norte (CBO, 2021). Além disso, a análise manual de grandes quantidades de imagens está sujeita à fadiga, variações subjetivas e limitações de tempo (NGUYEN et al., 2019). Nesse cenário, sistemas automáticos baseados em técnicas de aprendizado profundo têm-se mostrado promissores para apoiar o rastreio, diagnóstico e monitoramento do glaucoma, oferecendo rapidez, precisão e reprodutibilidade.

Esta pesquisa visa analisar, propor e desenvolver um método baseado em aprendizado profundo para classificação do estágio do glaucoma, utilizando retinografias e OCTs como entradas, não somente para facilitar o diagnóstico precoce, mas também para contribuir na identificação de casos avançados que demandam intervenção imediata. O método investigará estratégias de otimização de modelos, arquiteturas multimodais e mecanismos de combinação de características para explorar as informações disponibilizadas por cada modalidade de imagem.

1.1 Hipóteses de Pesquisa

O diagnóstico de doenças oftalmológicas é realizado utilizando-se de análise de campo visual, dados do indivíduo e exames de imagem. Estes exames permitem a visualização de biomarcadores que possibilitam detectar alguma condição e determinar o seu estágio de evolução. Os exames de retinografia e tomografia de coerência óptica são frequentemente utilizados para avaliar as estruturas retinianas, sendo ferramentas de grande importância no diagnóstico de glaucoma, retinopatia diabética, degeneração da mácula e outras condições. Enquanto a retinografia fornece a visualização do disco óptico, da escavação e dos vasos sanguíneos da retina, a tomografia fornece cortes de diferentes áreas da retina, sendo possível analisar e mensurar a espessura de camadas retinianas,

como a de fibras nervosas e a de células ganglionares. Sendo exames complementares, é recomendável a utilização de ambos para um diagnóstico mais preciso. Baseando-se nessas observações, foram consideradas as seguintes hipóteses para o trabalho.

Hipótese 1: A utilização de mais de uma modalidade de imagem médica oftalmológica pode elevar a assertividade do diagnóstico do estágio de doenças por possibilitar aos modelos a detecção de diferentes biomarcadores, o que pode levar os modelos a realizar predições mais precisas.

As modalidades médicas oftalmológicas existentes proporcionam a visualização de diferentes áreas dos olhos, sendo utilizadas por especialistas para verificar a existência de anormalidades características de condições que afetam a acuidade visual. Estas anormalidades são denominadas biomarcadores, e podem estar associadas a uma ou mais condições, como o aumento da escavação do nervo óptico, que pode ser associado ao glaucoma, e é visível em exames como a retinografia (WU et al., 2022; TAN; WONG, 2023), e o afinamento das camadas da retina, sendo uma lesão frequentemente associada à doenças oculares, que pode ser visualizada em tomografias de coerência óptica (ELGAFI et al., 2022). A disponibilidade de diferentes modalidades permite aos especialistas análises mais precisas para determinar a existência de uma condição e classificar o seu estágio de severidade. É suposto ser possível propor e ajustar modelos que utilizem mais de uma modalidade para detecção e classificação de doenças oftalmológicas, e que estes aprendam a encontrar características visuais correspondentes aos biomarcadores presentes em cada modalidade. Deste modo, é esperado que o uso de diferentes modalidades possibilite alcançar classificações mais precisas.

Hipótese 2: Considerando que a retinografia e a tomografia de coerência óptica são imagens que apresentam características visuais muito diferentes entre si, o emprego de técnicas de *ensemble* é a estratégia de fusão de características mais adequada para a tarefa de classificação utilizando um modelo multimodal.

Modelos multimodais devem possuir a capacidade de processar e extrair informações de mais de um tipo de dado de entrada. Estes modelos podem ser empregados em diferentes tarefas, sendo uma delas a classificação de imagens. Para possibilitar o desenvolvimento de modelos com duas ou mais entradas e uma saída, estratégias de fusão de características podem ser utilizadas para que estes modelos realizem a classificação automática de imagens. Estas estratégias se baseiam na fusão (combinação) de dados em algum ponto da arquitetura do modelo, sendo mais usual a fusão dos dados antes da entrada no modelo (*input-level fusion*), fusão após as camadas de extração de características (*early fusion*) e a fusão das probabilidades geradas por diferentes modelos (*ensemble* ou *late fusion*) (BOULAHIA et al., 2021; PAWŁOWSKI; WRÓBLEWSKA; SYSKO-ROMAŃCZUK, 2023). Logo, a escolha da melhor estratégia deve considerar a natureza dos dados disponíveis.

Hipótese 3: Modelos de aprendizado profundo desenvolvidos para detecção de doenças oculares em imagens de tomografia de coerência óptica podem alcançar maior desempenho diagnóstico quando treinados para focar em regiões de interesse restritas às camadas retinianas mais relevantes para cada condição específica.

O glaucoma provoca alterações estruturais e perda progressiva de células ganglionares e suas fibras nervosas, resultando em afinamentos detectáveis em cortes transversais da retina (ZHANG et al., 2016). Assim, restringir a análise computacional a estas camadas pode direcionar a capacidade discriminativa do modelo para regiões de maior relevância clínica, (CHEN et al., 2018). A delimitação da análise às camadas acima da coróide — como a camada de fibras nervosas da retina e a camada de células ganglionares — tende a potencializar a sensibilidade dos modelos, dado que as alterações estruturais iniciais desta doença ocorrem preferencialmente nestas regiões.

1.2 Objetivo Geral

O objetivo geral desta tese é analisar e desenvolver um modelo multimodal de classificação de glaucoma baseado em técnicas de aprendizado profundo, utilizando como entrada duas modalidades de imagem médica — retinografia e tomografia de coerência óptica. Além disso, busca-se investigar como a detecção automática da região de interesse em imagens OCT pode contribuir para o aumento da acurácia de modelos aplicados à classificação de estágio de glaucoma, ampliando assim o potencial clínico e diagnóstico do sistema proposto.

1.3 Objetivos Específicos

Para alcançar o objetivo geral deste trabalho, faz-se necessário atingir os seguintes objetivos específicos:

- Desenvolver e avaliar arquiteturas multimodais formadas por diferentes redes neurais convolucionais para classificação do estágio de doenças oftalmológicas;
- Avaliar modelos para extração de características visuais de retinografias e tomografias de coerência óptica, analisando e comparando o desempenho de modelos unimodais e multimodais, buscando determinar quais biomarcadores foram detectados e tiveram mais relevância na classificação final;
- Avaliar estratégias que permitam a fusão de características extraídas de cada uma das modalidades, realizando a comparação entre os métodos mais utilizados na literatura;

- Utilizar mecanismos de explicabilidade que permitam entender qual região das modalidades foi determinante para a classificação.

1.4 Contribuições

Destacam-se como principais contribuições oriundas do método desenvolvido nesta tese:

1. Otimização de modelos para classificação do estágio de severidade do glaucoma, com foco na seleção dos melhores extratores de características para cada modalidade de imagem — retinografia e OCT — e na combinação eficaz dessas representações.
2. Avaliação do impacto das modalidades de imagens de entrada para o resultado do modelo;
3. Avaliação de estratégia de fusões de características em modelos multimodais sobre imagens de OCT e retinografia;
4. Avaliação do impacto da pré-segmentação de estruturas do olho presentes nas imagens como forma de melhoria do diagnóstico.

1.5 Organização do Trabalho

Os demais capítulos deste trabalho foram organizados em:

- O Capítulo 2 trata dos conceitos fundamentais necessários para a construção desta pesquisa.
- O Capítulo 3 apresenta um resumo dos trabalhos relacionados à classificação de glaucoma usando modelos multimodais.
- O Capítulo 4 descreve todas as etapas do método proposto usado para a classificação multimodal de estágio de glaucoma.
- O capítulo 5 apresenta e discute os resultados alcançados;
- Finalmente, o Capítulo 6 apresenta as considerações finais e sugestões de trabalhos futuros.

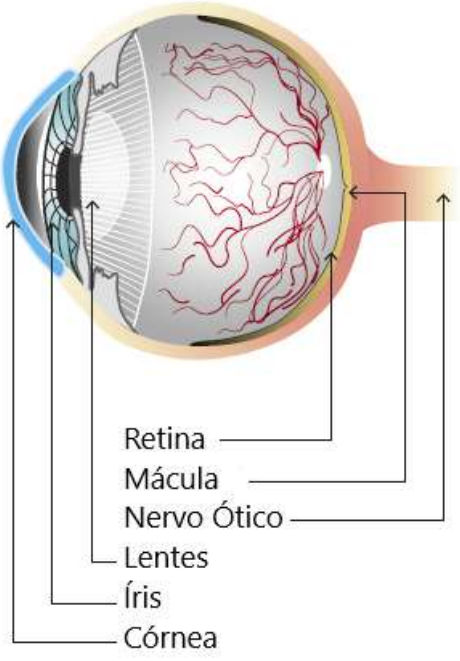
2 Fundamentação Teórica

Neste capítulo são apresentados os tópicos necessários para a compreensão das técnicas usadas na elaboração do método proposto. As seções a seguir abordam conceitos sobre glaucoma, modalidades de imagens médicas oftalmológicas, aprendizado profundo e redes neurais convolucionais e combinação de características visuais.

2.1 Olho Humano, Patologias e Exames de Imagem

O sistema visual abrange os olhos, os nervos ópticos e as vias de acesso entre diferentes estruturas do cérebro. Estruturas presentes na região frontal do olho, córnea e cristalino, recebem a luz que entra no olho e vai para a região posterior, em que se encontra a retina. Na retina, a luz é convertida em impulsos nervosos que viajam através do nervo óptico em direção a uma parte específica do cérebro conhecida como córtex visual (WHO, 2019). Esses impulsos são então transmitidos para outras partes do cérebro, onde eles se integram com outros sinais (como audição ou memória) para permitir que uma pessoa compreenda o ambiente circundante e responda de acordo. A Figura 1 apresenta as principais partes da anatomia do olho humano.

Figura 1 – Anatomia do Olho Humano.



Fonte: (WHO, 2019).

Diversas condições podem afetar as funções visuais, como a acuidade visual, que se refere à habilidade de se enxergar detalhes, independentemente da distância do objeto.

Fatores como a idade, disposição genética, estilo de vida, infecções e outros fatores relacionados à saúde contribuem para o surgimento e desenvolvimento de condições como glaucoma. O glaucoma é apontado como a principal causa de cegueira irreversível, com uma estimativa de aproximadamente 120 milhões de pessoas afetadas pela doença em 2040 (THAM et al., 2014).

2.1.1 Glaucoma

O glaucoma é apontado como a segunda principal causa de cegueira, com números inferiores à catarata. Os danos visuais provocados por glaucoma são irreversíveis, enquanto a catarata pode ser reversível cirurgicamente (WHO, 2019). É uma doença neurodegenerativa que possui diversas origens e que afeta progressivamente o nervo óptico e o campo visual. O glaucoma é uma condição que tem como principal fator de risco uma elevada pressão intraocular (PIO), que provoca danos progressivos ao nervo óptico, levando à diminuição da visão periférica, que pode evoluir para uma perda total da visão. A perda de visão progressiva provocada pelo glaucoma é ilustrada na Figura 2 (SOUZA et al., 2023).

Figura 2 – Exemplos representativos dos estágios de evolução do glaucoma.

Fonte: (WU et al., 2022).

A doença pode ser classificada como adquirida ou congênita, com pressão intraocular normal ou elevada, de ângulo aberto ou fechado. Neste último caso, a drenagem do líquido que preenche e é responsável pela lubrificação do olho humano, o

humor aquoso, é diminuída devido ao fechamento do ângulo formado pelo íris e pela córnea, levando ao aumento da pressão (SBG, 2012).

Além da pressão elevada, outros fatores de risco, como idade mais avançada, histórico familiar e raça, devem ser considerados (SBG, 2009). Apesar de ser irreversível, a perda da visão pode ser evitada, caso procedimentos médicos sejam realizados em estágios iniciais da doença para evitar o seu progresso (WHO, 2019). No entanto, como o glaucoma é assintomático nos estágios iniciais, o diagnóstico precoce é dificultado, sendo que a doença é geralmente diagnosticada quando os danos ao nervo óptico e a degeneração das células retinianas já provocaram danos à visão (SARHAN; ROKNE; ALHAJJ, 2019).

2.1.2 Modalidades de Imagens Oftalmológicas

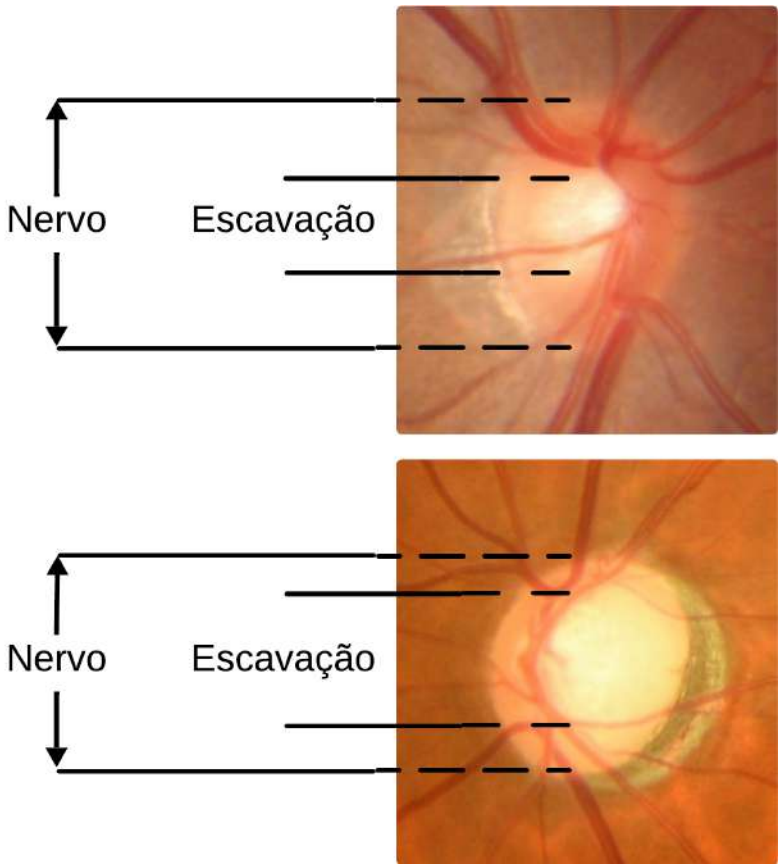
Diferentes modalidades de imagens médicas são usadas para identificar o estágio de doenças juntamente com dados de triagem, como idade, doenças crônicas e histórico familiar. As modalidades existentes permitem aos especialistas identificar e quantificar biomarcadores que caracterizam cada fase.

A retinografia ou fotografia colorida do fundo de olho é uma técnica rápida, não invasiva, bem tolerada e amplamente disponível que produz imagens de retina de alta qualidade. Utilizando esta modalidade, os especialistas podem analisar a região do nervo óptico, detectando aumento de sua escavação, sendo uma característica frequente em pacientes com glaucoma. A escavação é a área central do disco óptico, que tem uma tonalidade de cor mais clara que o disco. Alterações podem ser identificadas pela detecção do aumento da escavação do nervo óptico (Figura 3). Para verificar o aumento, podem ser calculadas as razões horizontal, vertical e de área entre o nervo e a escavação.

O aumento da pressão intraocular prejudica a circulação sanguínea e a nutrição das células nervosas do nervo óptico, sendo uma das causas do aumento da escavação. O resultado do processo provocado pelo aumento da pressão é a morte das células do nervo óptico (WANG et al., 2023). Esta modalidade também permite a visualização dos vasos sanguíneos da retina. Segundo estudos, a doença pode provocar anormalidades nestes vasos, como redução notável do diâmetro vascular da retina, diminuição da densidade vascular da retina, perturbações das moléculas vasculares da retina, diminuição da autorregulação do fluxo sanguíneo, disfunção das arteríolas da retina e casos de oclusão das veias da retina (WANG et al., 2023). Um estudo sugere que os modelos de detecção baseados em aprendizado profundo, não utilizam apenas a região do disco óptico para classificação das imagens, e sim toda a região da retina (HEMELINGS et al., 2021), em que se encontram os vasos sanguíneos.

Além da retinografia, existe a retinografia de campo ultra amplo (UWF, do inglês

Figura 3 – Amostras de imagens classificadas como sem glaucoma (acima) e com glaucoma moderado ou avançado (abaixo).

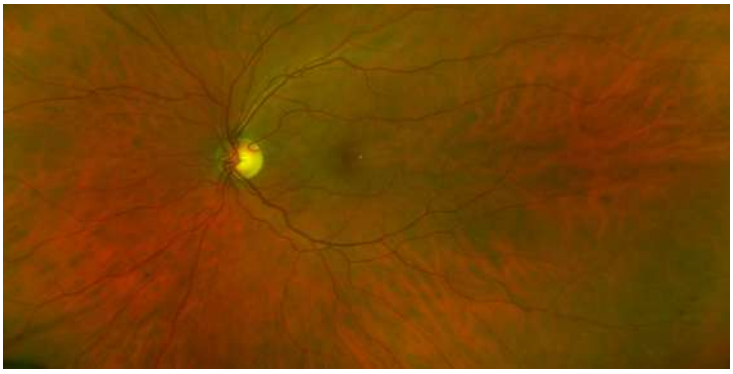


Fonte: (WU et al., 2022).

Ultra-widefield fundus photography) (Figura 4). Esta modalidade baseia-se numa tecnologia que capta até 200° do campo numa única imagem, em comparação com a fotografia de fundo de olho padrão, que capta entre 30 e 55° do campo numa única imagem. A imagem UWF tem várias vantagens adicionais, como a redução do tempo de aquisição da imagem. A UWF tem o potencial de identificar cerca de 17% mais biomarcadores do que a fotografia de fundo de olho de campo único (SOLIMAN et al., 2012).

Uma modalidade derivada da retinografia é a angiografia fluoresceínica (AF). Este é um exame complementar utilizado para ultrapassar as limitações da retinografia na detecção de lesões da retina, como aneurismas e hemorragias (HERVELLA et al., 2022). Para obter este exame, é necessário injetar nos pacientes um agente de contraste sanguíneo denominado fluoresceína. Como resultado, é possível destacar os vasos sanguíneos e as lesões presentes na retina, permitindo uma análise mais detalhada e facilitando a visualização de biomarcadores. As AF tradicionais são usadas para analisar 30° - 50° da retina, mas a AF de campo amplo estende o campo de visão até 200° (RABIOLO et al., 2017). A Figura 5 apresenta amostras de retinografia e angiografia, com

Figura 4 – Retinografia de campo amplo.



Fonte: (LUCENTE et al., 2023).

uma lesão destacada, que possui maior contraste na angiografia.

Figura 5 – Angiografia Fluoresceínica.

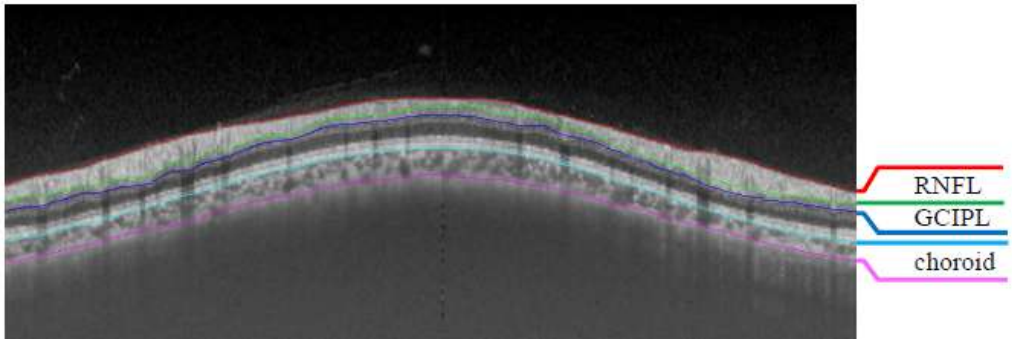
Fonte: (ALIPOUR; RABBANI; AKHLAGHI, 2014).

A tomografia de coerência óptica (OCT) é uma técnica de imagem que não requer contato e utiliza interferometria de baixa coerência para medir a luz refletida de diferentes camadas da retina e do nervo óptico, fornecendo medições quantitativas das estruturas do segmento posterior do olho, notadamente da mácula, limites da cabeça do nervo óptico e escavação, comparando-os a um banco de dados normativo para auxiliar o oftalmologista no diagnóstico e acompanhamento de doenças da retina. É uma modalidade de imagem essencial para avaliar o glaucoma. Esta modalidade permite detectar afinamento em camadas da retina, como a camada de fibras nervosas da retina (RNFL) e a camada plexiforme de células ganglionares (GCIPL), além de possibilitar a visualização de edemas maculares e desorganização das camadas internas da retina. O afinamento das camadas da retina é uma alteração presente em pacientes com glaucoma (WU et al., 2022), degeneração macular e retinopatia diabética (ELGAFI et al., 2022) e está relacionado à morte de células retinianas e ganglionares (WANG et al., 2023).

A Figura 6 apresenta a marcação das camadas RNFL, GCIPL e do coroide em uma amostra de OCT e a Figura 7 apresenta amostras de OCTs de pacientes com edema macular diabético. Apesar do aumento do uso da OCT para diagnóstico, a experiência

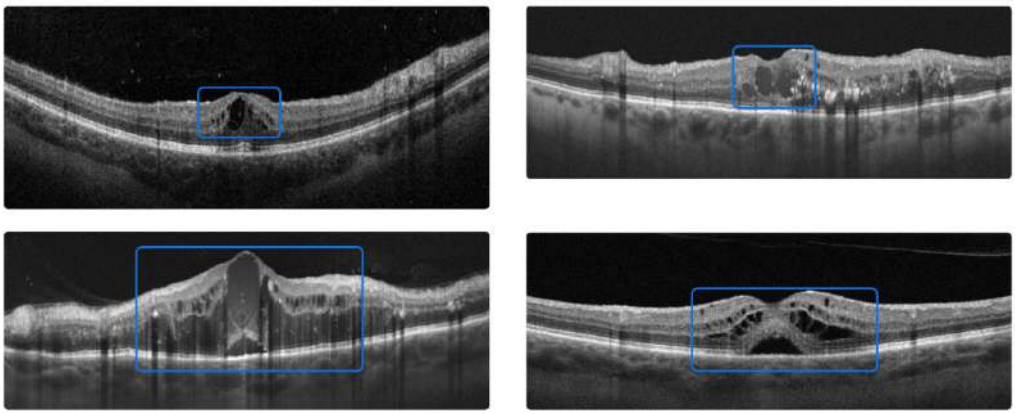
humana na interpretação das imagens OCT ainda é limitada, cara e de difícil transferência de aprendizado entre especialistas, limitando a viabilidade da adoção generalizada nos esforços de triagem (RUIA; TRIPATHY, 2021).

Figura 6 – Camadas da retina em uma imagem OCT, com destaque para a camada de fibras nervosas da retina (RNFL), a camada de células ganglionares (GCIPL) e a coróide (choroid).



Fonte: (FANG et al., 2022).

Figura 7 – Amostras de imagens OCT de indivíduos com edema macular diabético, com as regiões de edema destacadas.

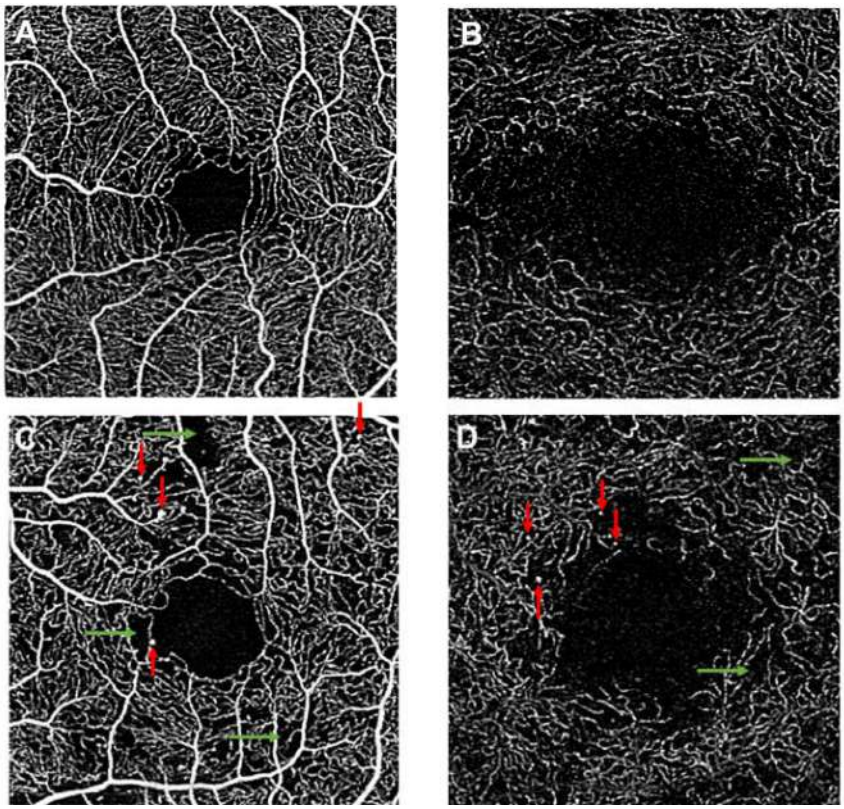


Fonte: (KULYABIN et al., 2024).

A angiografia por tomografia de coerência ótica (OCTA, do inglês *Optical Coherence Tomography Angiography*) permite a investigação não invasiva e individual das camadas vasculares da retina para delinear com precisão áreas não vascularizadas e determinar a causa da redução da visão em doentes com retinopatia diabética sem edema central. Permite a visualização dos 3 plexos capilares da retina (plexos capilares profundos, superficiais e médios) e do plexo coroide sem utilização de corante ou contraste. Gera imagens nos glóbulos vermelhos à medida que se deslocam através dos vasos da retina. Identifica aspectos importantes da retinopatia, como áreas de alto fluxo e isquemia (não perfusão), microaneurismas e neovascularização. No entanto, a maioria dos equipamentos de OCTA não permite a obtenção de imagens da retina periférica

(SUCIU et al., 2020). A Figura 8 apresenta amostras de OCTAs de um olho saudável, no topo, e de um olho de um indivíduo com retinopatia.

Figura 8 – OCTA de um indivíduo saudável no topo e uma OCTA de um indivíduo com diabetes, mostrando microaneurismas (setas vermelhas) e não perfusão capilar (setas verdes).



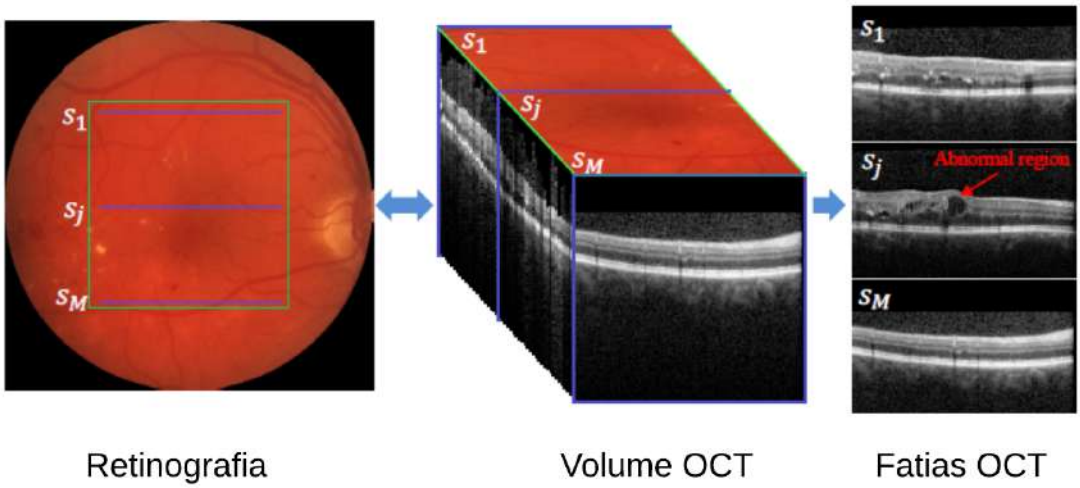
Fonte: (CHUA et al., 2020).

2.2 Relação entre retinografia e tomografia de coerência óptica

A tomografia de coerência óptica é um exame de imagem que realiza varreduras em regiões oculares, permitindo a visualização em alta resolução das diferentes camadas da retina. Usualmente, esse exame é centralizado na mácula, estrutura anatômica situada na região central da retina (WANG et al., 2024). Por sua vez, a mácula também pode ser identificada em retinografias, aparecendo como uma região de coloração mais escura próxima ao nervo óptico (Figura 9). Enquanto a retinografia fornece uma visão bidimensional da superfície retiniana, ressaltando a vascularização e a anatomia geral do fundo de olho, o OCT possibilita a análise seccional e em profundidade, revelando alterações estruturais que não são visíveis no exame fotográfico. A região central da mácula, denominada fóvea, apresenta-se de forma característica em ambos os exames: na retinografia pode ser delimitada topograficamente, enquanto no OCT é possível

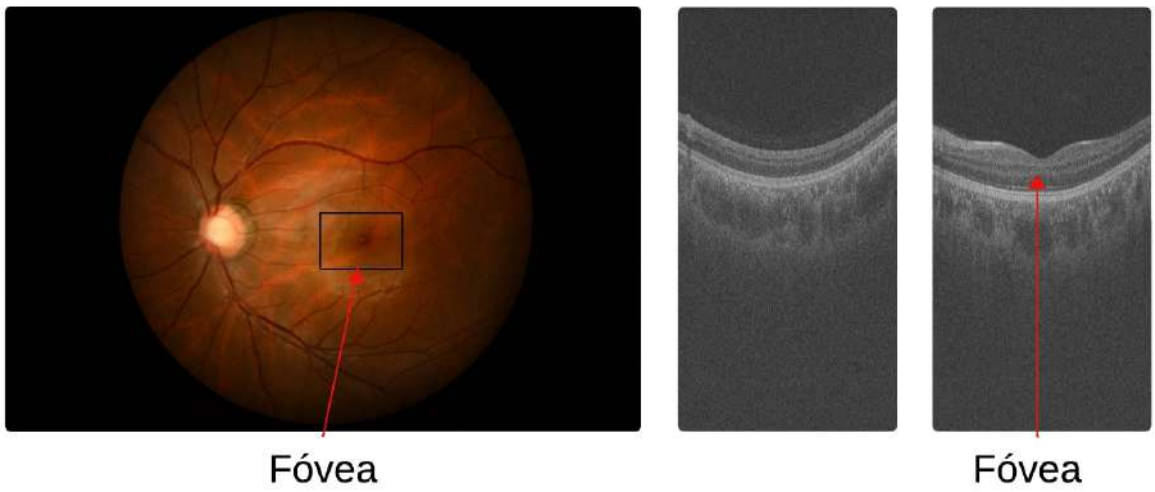
observar cortes com curvatura específica dessa área (Figura 10). Ressaltar essa relação entre as duas modalidades é importante, pois, além de serem amplamente utilizadas de forma complementar na prática clínica, correspondem exatamente aos tipos de imagem empregados nesta pesquisa.

Figura 9 – Relação entre uma retinografia e o correspondente volume OCT.



(WANG et al., 2024)

Figura 10 – Localização da fóvea em uma retinografia. Fatia OCT sem (centro) e com fóvea (à direita).



Fonte: imagem elaborada pelo autor.

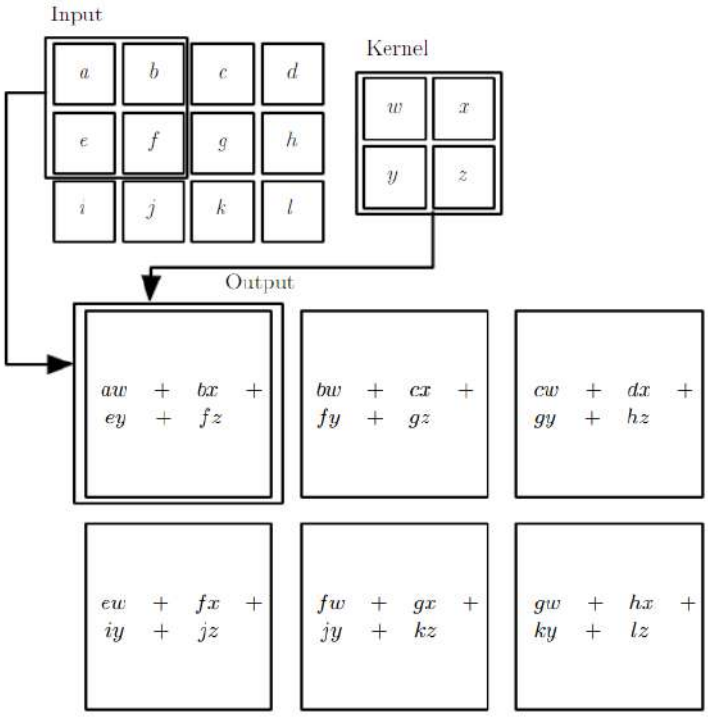
2.3 Redes Neurais Convolucionais

Redes Neurais Convolucionais (CNN, do inglês *Convolutional Neural Network*) são uma classe de redes neurais profundas usadas com sucesso em aplicações de processamento e reconhecimento de imagens. São formadas por diferentes tipos de

camadas, destacando-se as camadas convolucionais, *pooling* e totalmente conectadas, que possuem uma função específica na propagação do sinal de entrada.

Após a entrada na rede, as imagens passam por camadas que realizam a operação de convolução por meio de um conjunto de filtros (*kernels*) de pequenas dimensões que se movem a um passo (*stride*) por toda a imagem (*input*) (Figura 11). Cada filtro é formado por um conjunto de pesos ajustados durante o processo de treinamento, para ocorrer o aprendizado para extração de características em diferentes regiões da imagem (ARAÚJO et al., 2017). O resultado do somatório dos produtos ponto a ponto passa por uma função de ativação, por exemplo, a função ReLU (*Rectified Linear Units*), calculada pela Equação 2.1. O resultado do processo de convolução utilizando todos os filtros é um mapa de características visuais (*feature maps*) (GOODFELLOW; BENGIO; COURVILLE, 2016).

Figura 11 – Exemplo de operação de convolução, responsável por extrair padrões locais da imagem por meio de filtros deslizantes.



Fonte: (GOODFELLOW; BENGIO; COURVILLE, 2016).

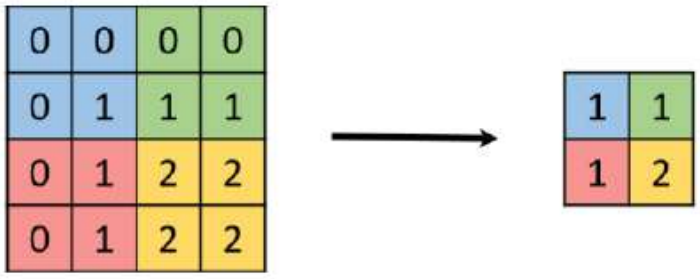
$$f(x) = max(0, x) \tag{2.1}$$

onde x consiste no sinal de entrada.

Além da camada convolucional, é comum existir uma camada de *pooling*. Esta camada é utilizada para reduzir a dimensão espacial dos mapas de características, reduzindo desta forma o custo computacional para treinamento da rede e diminuindo a

possibilidade de *overfitting* (ARAÚJO et al., 2017). Na operação de *pooling*, valores pertencentes a uma determinada região do mapa de características são substituídos por um único valor, que pode ser calculado pela média dos valores (*average pooling*), pelo valor máximo (*max pooling*), apresentado na Figura 12, dentre outros (GOODFELLOW; BENGIO; COURVILLE, 2016). O conjunto de camadas convolucionais e camadas de *pooling* forma o extrator de características de uma CNN, apresentado na Figura 13.

Figura 12 – Exemplo de operação de *Max Pooling*, que reduz a dimensão espacial mantendo as características mais relevantes.



Fonte: (ARAÚJO et al., 2017).

Após as camadas que formam o extrator, existe uma camada que realiza a operação de *flatten* (vetorização) dos mapas de características. Após este processo, as características são utilizadas como entrada para um classificador, geralmente formado por camadas densas (que possuem conexões entre todos os seus neurônios). Em termos matemáticos, um neurônio de uma camada densa pode ser descrito pelas Equações 2.2 e 2.3.

$$h_j = \sum_{i=1}^{m} w_{ij} x_i \tag{2.2}$$

$$y_j = \varphi(h_j + b_j) \tag{2.3}$$

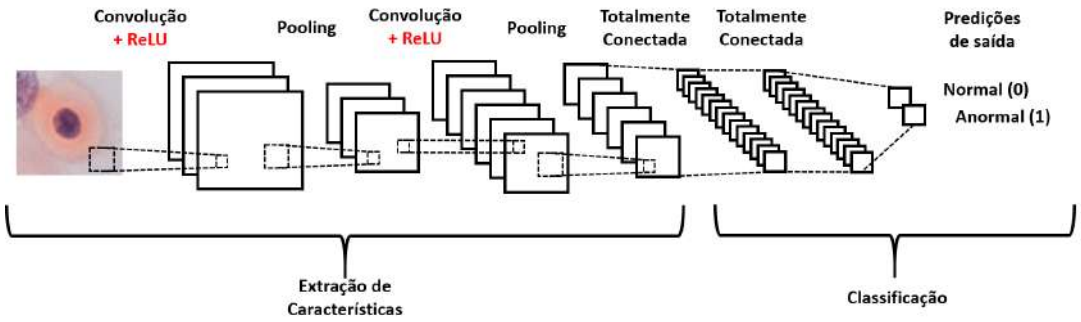
onde x_i são os m sinais de entrada, w_{ij} são os pesos sinápticos do neurônio j, e b_j corresponde ao *bias*, responsável por realizar o deslocamento da função de ativação definida por φ (ARAÚJO et al., 2017). A última camada do classificador possui uma quantidade de neurônios equivalente ao número de classes da tarefa. Essa camada tem como saída uma distribuição de probabilidades calculada pela função *softmax* (BISHOP, 2006) (Equação 2.4).

$$\sigma(Z)_i = \frac{e^{zi}}{\sum_{j=1}^{K} e^{zj}} \tag{2.4}$$

onde z é um vetor de números reais normalizado em uma distribuição de K probabilidades. Para obter-se o resultado da previsão da classificação, é necessário verificar qual

neurônio tem como resultado a maior probabilidade da distribuição.

Figura 13 – Estrutura geral de uma Rede Neural Convolucional (CNN).



(ARAÚJO et al., 2017)

2.4 Aprendizado

Em uma rede neural convolucional, os pesos dos filtros das camadas convolucionais e das camadas totalmente conectadas podem ser inicializados com valores aleatórios ou com valores resultantes de um treinamento anterior, técnica denominada transferência de aprendizado, detalhada na Seção 2.6. É necessário que esses pesos sejam ajustados para que a tarefa de classificação ou segmentação tenha uma precisão satisfatória.

O método mais utilizado para treinamento de redes neurais é o algoritmo de *backpropagation* (GOODFELLOW; BENGIO; COURVILLE, 2016). Após a inicialização dos pesos, a rede recebe os dados de entrada e realiza o processo de propagação, por meio das operações de convolução, função de ativação e *pooling*. O resultado desse processo é a distribuição de probabilidades na camada de saída da rede. Em seguida, é calculado o erro obtido por meio de uma função de *loss*, ou função de custo. Em seguida, é calculado o vetor gradiente do erro para ser possível determinar a direção do menor valor da função de erro para ajuste apropriado dos pesos, a partir da última camada em direção à primeira (retropropagação). Deste modo, espera-se que o erro obtido a cada iteração de treinamento (época) seja menor até que atinja um valor que satisfaça uma condição de parada do treinamento (ARAÚJO et al., 2017).

Para acelerar a velocidade de aprendizagem, são utilizados otimizadores, como SGD (do inglês, Stochastic Gradient Descent) e Adam. Estas técnicas de otimização são responsáveis pelo ajuste dos pesos das redes e dependem de um parâmetro que determina a velocidade de aprendizado e é fundamental para a rede neural ser capaz de executar a tarefa para a qual foi treinada: a taxa de aprendizado (LR, do inglês *Learning Rate*). Se o valor desta taxa for muito baixo, a rede irá necessitar de muitas épocas para ocorrer a conversão para uma solução ótima. Mas se for muito alto, pode levar a uma oscilação do desempenho ao longo das épocas de treinamento e a um desempenho final

inferior ao esperado. A Equação 2.5 apresenta a relação da taxa de aprendizado e da atualização de pesos, quando é empregado o otimizador SGD:

$$weight_{t+1} = weight_t - lr * \frac{\partial error}{\partial weight_t} \tag{2.5}$$

onde $weight_{t+1}$ é o valor dos pesos que serão utilizados na próxima época de treinamento, lr é a taxa de aprendizado, $error$ é a função de custo utilizada para cálculo do erro e $weight$ é valor atual dos pesos, equação adaptada de (GOODFELLOW; BENGIO; COURVILLE, 2016).

2.5 Fusão de Características

Modelos multimodais para classificação são modelos baseados em aprendizado de máquina, que recebem mais de um dado como entrada, podendo ser imagens, textos, dados clínicos. Pela necessidade destes modelos lidarem com duas ou mais modalidades de imagem simultaneamente, em algum momento, alguma estratégia deve ser utilizada para combinar as características extraídas de cada imagem.

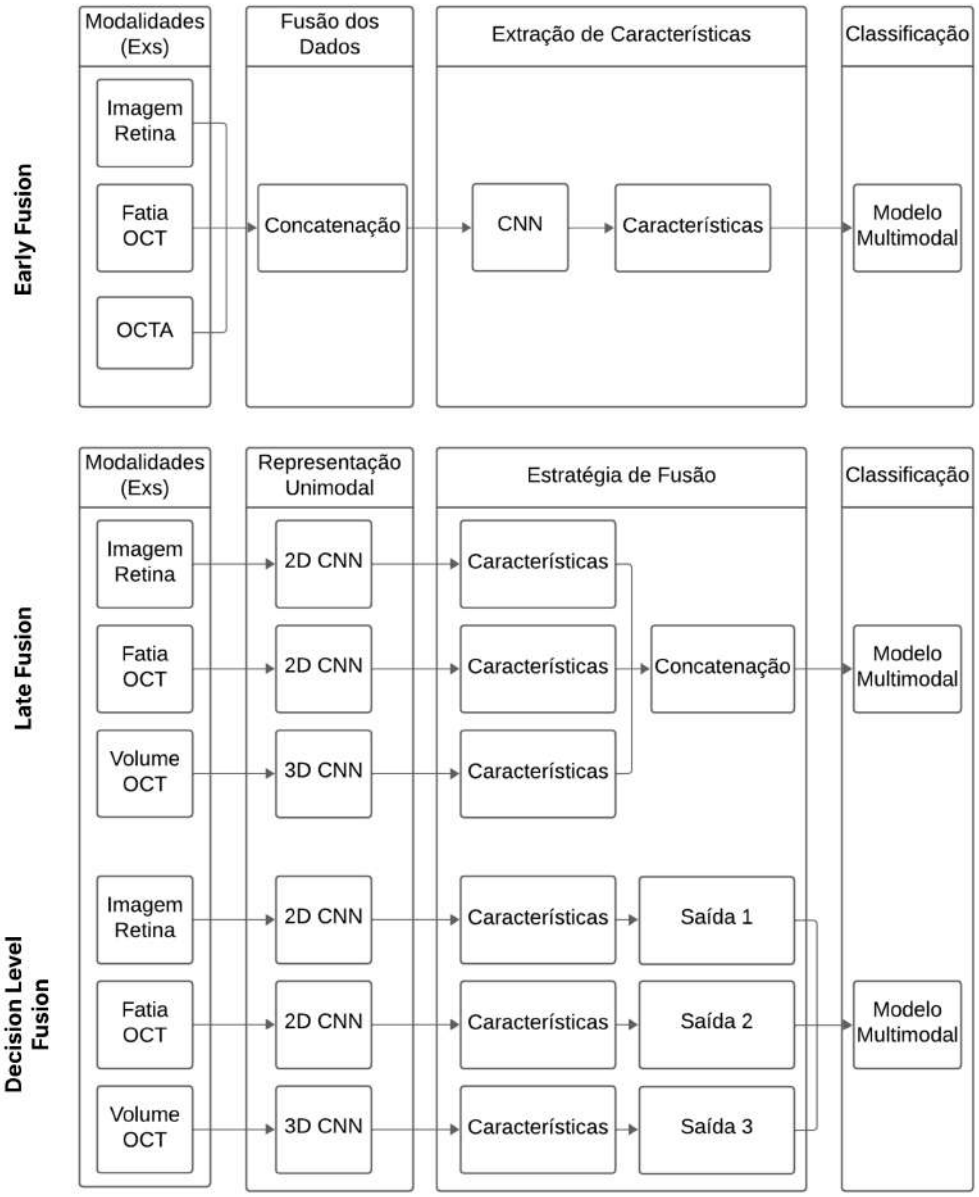
A Figura 14 apresenta uma representação de três estratégias frequentemente utilizadas: *Early Fusion* (concatenação dos canais das imagens antes da camada de entrada dos modelos), *Late Fusion* (concatenação dos mapas de características extraídos dos *backbones*) e *Decision Level Fusion* (estratégias de *ensemble* para combinar as saídas dos classificadores de cada modelo (BOULAHIA et al., 2021), (PAWŁOWSKI; WRÓBLEWSKA; SYSKO-ROMAŃCZUK, 2023).

Na estratégia *early fusion*, as modalidades utilizadas são concatenadas no eixo de referência ao número de canais, gerando uma nova imagem que combina as duas modalidades e será utilizada como entrada. Esta estratégia requer um modelo que contenha uma etapa para realizar a extração de características.

Na estratégia *late fusion*, as modalidades alimentam o modelo separadamente, exigindo um nível de extração de características para cada modalidade. Os mapas extraídos por cada nível são então concatenados antes de serem utilizados para classificação.

Nas estratégias de ensemble, cada modalidade alimenta um modelo com um extrator e um classificador. O resultado da classificação é calculado utilizando as probabilidades geradas por cada modelo. Um dos métodos mais utilizados é o Ensemble médio, no qual é calculada uma média simples das probabilidades geradas por cada modelo (BOULAHIA et al., 2021). O emprego das estratégias de late fusion e ensemble foi avaliado para classificação multimodal na Seção 4.3.

Figura 14 – Estratégias para fusão de características.



Adaptada de (PAWŁOWSKI; WRÓBLEWSKA; SYSKO-ROMAŃCZUK, 2023), (BOULAHIA et al., 2021)

2.6 Transferência de Aprendizado

Um problema recorrente em tarefas de classificação de imagens utilizando aprendizagem profunda. Deve-se ao fato de ser necessário um conjunto muito grande de imagens para ser possível realizar um treinamento eficiente de uma CNN (ARAÚJO et al., 2017). Logo, é raramente realizado o treinamento de uma CNN com inicialização aleatória de pesos, pois, em geral, conjuntos de imagens médicas não são suficientemente grandes. Uma das técnicas utilizadas para sanar esta limitação é a transferência de

aprendizado. Esta técnica consiste em se utilizar o que foi aprendido na classificação de um conjunto de dados para aumentar a capacidade de generalização em outro conjunto (GOODFELLOW; BENGIO; COURVILLE, 2016). Neste caso, o primeiro conjunto é muito maior que o segundo. Uma prática comum é a utilização dos pesos de uma CNN previamente treinada na base *ImageNet*, que possui mais de 1 milhão de imagens e 1000 classes (ARAÚJO et al., 2017).

Redes pré-treinadas podem ser utilizadas de diferentes formas em tarefas de classificação de imagens. Em geral, apenas as camadas que formam o extrator de características de uma CNN são aproveitadas de modelos pré-treinados, sendo que um novo classificador, ajustado à nova tarefa, é adicionado à rede. A estratégia *fine-tuning* consiste em dar continuidade ao treinamento de uma rede previamente treinada. É possível a realização de *fine-tuning* de todas as camadas de uma CNN ou somente das últimas camadas. Isto deve-se ao fato de que as primeiras camadas de uma rede convolucional possuem extratores mais genéricos, como detectores de bordas, enquanto as camadas mais profundas possuem detalhes específicos da base com a qual a rede foi previamente treinada (ARAÚJO et al., 2017). Também é possível utilizar uma CNN pré-treinada como um extrator de características. Neste caso, o classificador é removido, e as características extraídas são utilizadas como entrada para um classificador que é mais rapidamente treinado (ARAÚJO et al., 2017), por exemplo, *Random Forest* (HO, 1995).

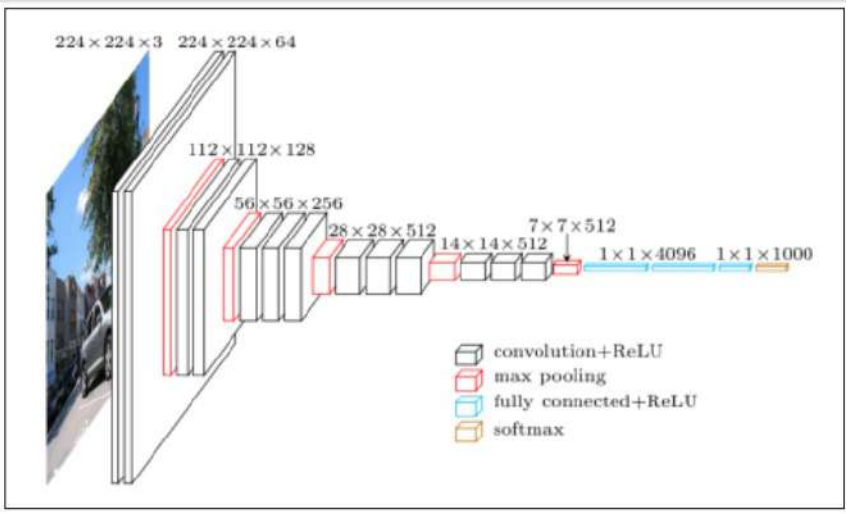
2.7 Arquiteturas de CNN

Existem atualmente diferentes arquiteturas de redes convolucionais. Os desenvolvedores têm um objetivo claramente definido, que é alcançar o melhor desempenho possível, sendo a meta classificar corretamente diferentes imagens. Anualmente é realizado um desafio denominado ImageNet Large-Scale Visual Recognition Challenge (ILSVRC) (RUSSAKOVSKY et al., 2015), que tem servido como teste para novas gerações de sistemas de classificação de imagem em larga escala (SIMONYAN; ZISSERMAN, 2014). A seguir, são apresentadas arquiteturas utilizadas para construção de modelos de classificação, seção 4.2 do capítulo de metodologia, utilizados nesta pesquisa.

A rede *Visual Geometry Group* (VGG), (SIMONYAN; ZISSERMAN, 2014), tem como principal característica o fato de utilizar filtros pequenos, o que possibilitou uma rede com muitas camadas. Um dos grandes problemas ao se aumentar o número de camadas é o desaparecimento do gradiente. A solução encontrada foi utilizar uma pilha de camadas convolucionais, onde são utilizados filtros com um campo receptivo muito pequeno (3x3). A utilização desses filtros reduz o custo computacional, sendo esta a grande contribuição da VGG. Além disso, esta rede utiliza cinco camadas de *Max Pooling* com uma janela de tamanho 2x2. Após as camadas convolucionais, há três camadas totalmente conectadas: as duas primeiras possuem 4096 neurônios cada e a última

possui 1000. A Figura 15 apresenta a arquitetura.

Figura 15 – Estrutura geral da rede VGG, composta por blocos convolucionais seguidos de camadas de pooling e totalmente conectadas.



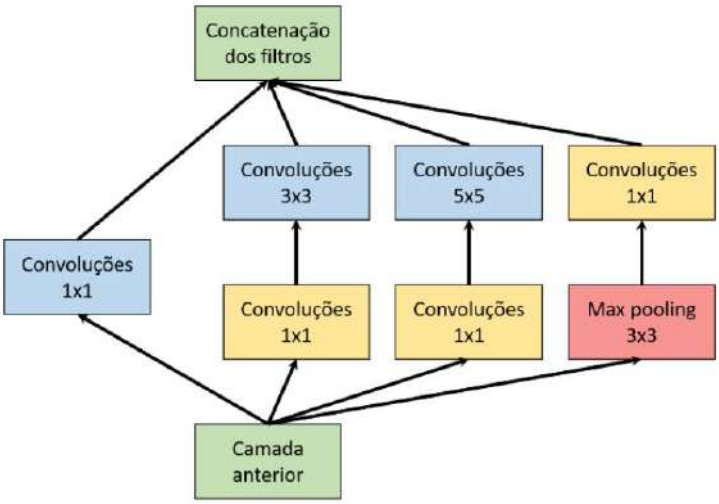
Fonte: (SIMONYAN; ZISSERMAN, 2014)

A rede Inception (SZEGEDY et al., 2015) é uma rede convolucional profunda que disputou o ILSVRC em 2014, atingindo o estado da arte naquele ano para classificação de imagens. Esta arquitetura introduziu a ideia de que as camadas não precisavam ser executadas sequencialmente, mas que podiam ser executadas paralelamente. Esta é a principal característica desse modelo (Figura 16). A grande vantagem dos módulos Inception é o uso de filtros 1x1 para reduzir o número de características dos blocos antes de realizar convoluções com filtros maiores (ARAÚJO et al., 2017). A utilização destes blocos permitiu o aumento da largura e profundidade da rede, sem aumento considerável do custo computacional.

As redes residuais (ResNet) (HE et al., 2016) trouxeram como grande contribuição conexões residuais entre as camadas de uma CNN. Em um bloco residual, um volume de entrada passa por uma sequência de convoluções e de ativações Relu. O resultado dessas operações é somado ao valor da entrada, como ilustrado na Figura 17. Em CNNs com conexões apenas *feedforward*, o volume de saída de um bloco convolucional é completamente diferente do volume de entrada x. Em relação ao modelo ResNet, o volume de saída H(x) é somente uma alteração do volume de entrada (ARAÚJO et al., 2017). Com estas conexões residuais foi possível implementar um modelo com até 152 camadas.

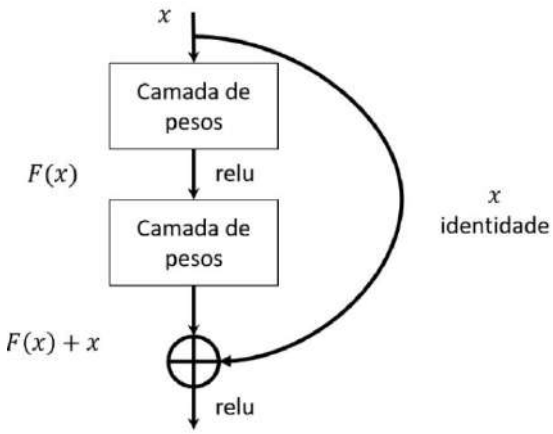
A DenseNet (*Dense Convolutional Network*) (HUANG; LIU; WEINBERGER, 2016) é uma rede convolucional que possui como principal característica conexões densas entre os blocos convolucionais. Redes convolucionais tradicionais conectam a saída de uma camada à entrada da próxima camada. Essa característica é uma das causas do

Figura 16 – Estrutura do módulo Inception, que realiza convoluções em múltiplas escalas (1×1, 3×3 e 5×5) e as concatena, permitindo a extração simultânea de características locais e globais com baixo custo computacional.



Fonte: (ARAÚJO et al., 2017)

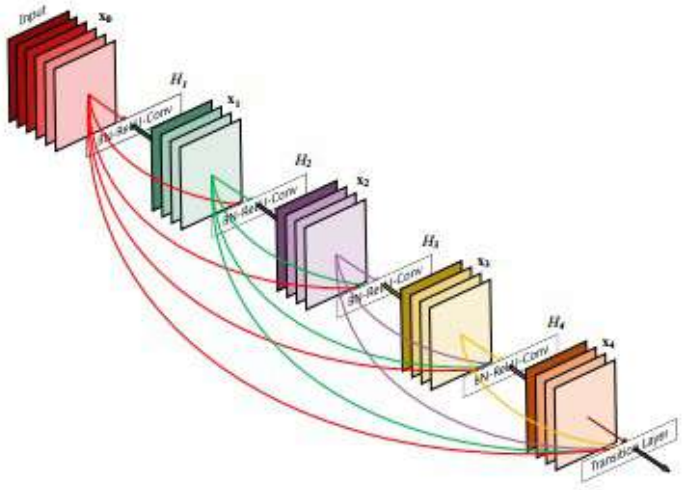
Figura 17 – Bloco básico da ResNet, com conexões de atalho (skip connections) que facilitam o treinamento de redes profundas.



Fonte: (ARAÚJO et al., 2017)

desaparecimento do gradiente em redes muito profundas. As redes residuais adicionaram uma conexão diferente das conexões até então utilizadas. A vantagem da conexão residual é que o gradiente pode fluir diretamente das últimas camadas para as primeiras. Já o modelo DenseNet propõe conexões diretas de qualquer camada para todas as camadas subsequentes. O modelo dessa conexão densa é ilustrado na Figura 18. Além dos blocos com conexões densas, existem camadas com conexões apenas *feedforward* que formam o classificador.

Figura 18 – Diagrama das conexões densas da DenseNet. Cada camada recebe como entrada os mapas de características de todas as anteriores, promovendo reutilização de informações.



Fonte: (HUANG; LIU; WEINBERGER, 2016)

2.8 Mecanismos de Atenção em Redes Neurais

Os mecanismos de atenção foram introduzidos para permitir que redes neurais aprendam a focar seletivamente em regiões ou canais mais informativos de uma representação de entrada. Na área de visão computacional, módulos de atenção são amplamente incorporados em arquiteturas convolucionais para melhorar a capacidade de extração de características discriminativas, com baixo custo computacional adicional. Neste trabalho, foram avaliados três mecanismos de atenção: *Squeeze-and-Excitation* (SE), *Spatial Attention* e o *Convolutional Block Attention Module* (CBAM) como uma tentativa de elevar a precisão de classificação dos modelos, seção 4.2.

2.8.1 Mecanismo SE (*Squeeze-and-Excitation*)

O bloco SE foi proposto por (HU; SHEN; SUN, 2018) para responder a canais de forma adaptativa, Figura 19. Para uma entrada $\mathbf{X} \in \mathbb{R}^{H \times W \times C}$, o bloco SE executa duas etapas: *squeeze* (compressão espacial) e *excitation* (recalibração de canais). Primeiramente, é realizada uma agregação global por *average pooling*:

$$\mathbf{z}_c = \frac{1}{HW} \sum_{i=1}^{H} \sum_{j=1}^{W} X_c(i, j)$$

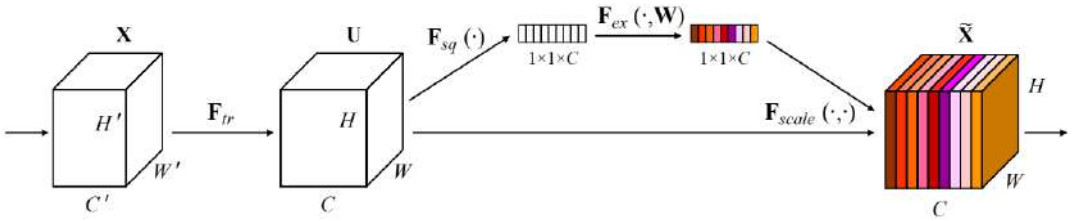
Em seguida, aplica-se uma operação de excitação, que consiste em duas camadas totalmente conectadas com não linearidade:

$$\mathbf{s} = \sigma\Big(\mathbf{W}_2\,\delta(\mathbf{W}_1\mathbf{z})\Big)$$

Onde δ é uma função ReLU e σ é uma função sigmoide. A saída é a reponderação dos canais de entrada:

$$\hat{\mathbf{X}}_c = s_c \cdot \mathbf{X}_c$$

Figura 19 – Mecanismo *Squeeze-and-Excitation*.



(HU; SHEN; SUN, 2018)

2.8.2 Mecanismo de Atenção Espacial

O módulo de Atenção Espacial (*Spatial Attention*), Figura 20, busca destacar regiões espaciais mais relevantes em cada mapa de ativação. Uma forma típica é computar uma máscara espacial a partir de operações de *average pooling* e *max pooling* ao longo da dimensão dos canais:

$$\mathbf{M}_s(\mathbf{X}) = \sigma\Big(f^{7\times7}([AvgPool(\mathbf{X}); MaxPool(\mathbf{X})])\Big)$$

Onde $f^{7\times7}$ representa uma convolução 7×7 e σ é uma função sigmoide. A entrada é então modulada por esta máscara:

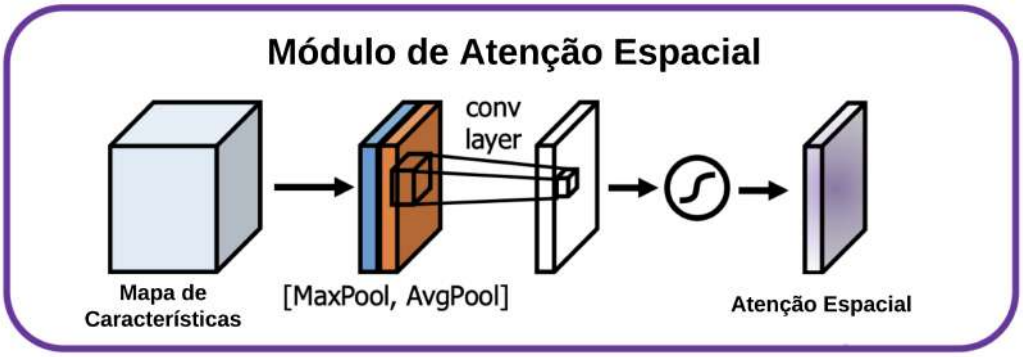
$$\hat{\mathbf{X}} = \mathbf{M}_s(\mathbf{X}) \odot \mathbf{X}$$

2.8.3 Mecanismo CBAM

O *Convolutional Block Attention Module* (CBAM) (WOO et al., 2018) combina atenção de canais e atenção espacial de forma sequencial, representado na Figura 21. Primeiro, aplica o bloco de atenção de canais como no SE, seguido da atenção espacial:

$$\mathbf{M}_c(\mathbf{X}) = \sigma\Big(MLP(AvgPool(\mathbf{X})) + MLP(MaxPool(\mathbf{X}))\Big)$$

Figura 20 – *Módulo de Atenção Espacial*.



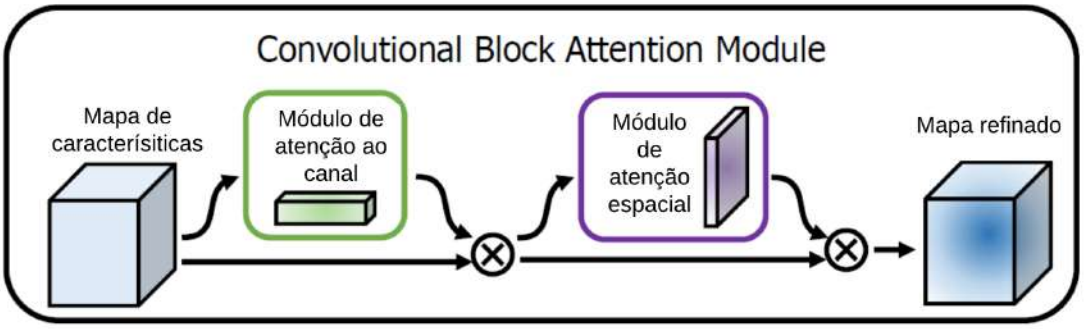
(WOO et al., 2018)

$$\mathbf{M}_s(\mathbf{X}) = \sigma\Big(f^{7\times7}([AvgPool(\mathbf{X}'); MaxPool(\mathbf{X}')])\Big)$$

onde $\mathbf{X}' = \mathbf{M}_c(\mathbf{X}) \odot \mathbf{X}$. A saída final do CBAM é dada por:

$$\hat{\mathbf{X}} = \mathbf{M}_s(\mathbf{X}') \odot \mathbf{X}'$$

Figura 21 – Mecanismo CBAM.



(WOO et al., 2018)

Esses mecanismos reforçam seletivamente regiões e canais mais informativos, contribuindo para uma representação mais robusta e discriminativa.

2.9 Considerações Finais

Neste capítulo foi apresentada a fundamentação teórica necessária para a compreensão das técnicas utilizadas e suas aplicações no método proposto. Foram apresentados conceitos sobre glaucoma, modalidades de imagens médicas oftalmológicas, aprendizado profundo e redes neurais convolucionais e fusão de características visuais.

3 Trabalhos Relacionados

Na literatura existem estudos relacionados que apresentam métodos desenvolvidos para detecção ou classificação de estágio de glaucoma usando uma ou mais modalidades de dados. Para esta pesquisa foram selecionados trabalhos relacionados que utilizaram dados multimodais para detecção de glaucoma em datasets privados e trabalhos que avaliaram seus métodos propostos para classificar o estágio de glaucoma utilizando imagens do dataset GAMMA (do inglês, *Glaucoma Grading from Multi-Modality Images*). As subseções a seguir apresentam resumos de trabalhos relacionados que apresentaram métodos para classificação de glaucoma usando dados multimodais.

O protocolo de busca considerou inicialmente trabalhos publicados a partir de 2018 que propuseram métodos para detecção de glaucoma baseados em múltiplas modalidades de imagem médica. Em seguida, foram priorizados os estudos que utilizaram o conjunto de dados GAMMA. Por fim, foram selecionados especificamente os trabalhos que abordaram a classificação de estágios do glaucoma, empregando as retinografias e OCTs disponibilizadas nesse dataset.

Em (AN et al., 2019), foram obtidos mapas de espessura e de desvio de camadas da retina de fibras nervosas e de células ganglionares. Além dos quatro mapas obtidos de OCTs, são utilizadas retinografias. Cada modalidade é utilizada para treinamento de modelos com a VGG19 como extrator de características, que são concatenadas e utilizadas como entrada para um classificador Random Forest, alcançando o valor AUC (do inglês, *Area Under the Curve*) de 0,96. Em outro trabalho, um modelo multimodal foi usado para realizar classificação binária de glaucoma combinando modelos formados por redes DenseNet para extração de características de volumes OCT e InceptionResNetV4 para extrair características de retinografias (MEHTA et al., 2021). As probabilidades geradas por cada modelo foram combinadas utilizando uma estratégia de ensemble com a probabilidade gerada por um modelo XGBoost, que recebeu como entrada dados dos pacientes, como idade, gênero, etnia e índice de massa corpórea, alcançando como melhor resultado um valor AUC de 0,97.

Em (MA et al., 2021) é descrito um método usado para classificar imagens de retinografia de pacientes em três classes: saudável, suspeito e em estágio inicial, baseando-se em 15 características clínicas, como a idade, PIO, espessura da camada RNFL, razões horizontal e vertical do disco óptico em relação à escavação. Foram realizados experimentos com diferentes classificadores, sendo o maior valor de AUC alcançado pelo classificador SVM (do inglês, *Support Vector Machine*), alcançando como melhor resultado um valor de 0,90. Em (XIONG et al., 2021) foi proposto um modelo

chamado FusionNet. Este modelo combina imagens de OCT com imagens obtidas de relatórios de avaliação de campo visual. Utiliza uma rede multinível customizada para cada uma das modalidades, com diferentes níveis formados por camadas convolucionais com módulos de atenção, seguida de um classificador que prevê as probabilidades de as imagens serem de um paciente com glaucoma. Como melhor resultado, o modelo alcançou um valor AUC de 0,95.

O estágio do glaucoma também foi objeto das seguintes pesquisas. Fang et al. (2021) descrevem uma arquitetura que combina características extraídas de imagens de retinografias e de OCTs. O objetivo é classificar as imagens sem glaucoma, em estágio inicial e em glaucoma em estágio moderado ou avançado, usando uma ResNet34.O melhor resultado alcançado foi um valor Kappa de 0,860, combinando as características extraídas de retinografias dos volumes da OCT e da região do disco óptico por meio de regressão ordinal. Da mesma forma, Li et al. (2022) apresentam um método que utiliza uma concatenação hierárquica de características realizada por uma rede alimentada por imagens de retinografia e imagens de OCT. Em cada nível da rede foram avaliadas variações da ResNet. Utilizando a técnica de concatenação proposta, foi possível aumentar o número total de características utilizadas para classificação, além de utilizar características de diferentes escalas. Os melhores resultados foram obtidos utilizando as ResNet50 e ResNet101 como extratores de características, alcançando os valores Kappa de 0,866 e 0,875, respectivamente.

O estudo apresentado em (CAI et al., 2022) descreve o *framework* COROLLA, um método de fusão de características com aprendizagem contrastiva supervisionada para classificação de glaucoma. O método proposto consiste em extrair mapas de espessura das camadas de retina presentes em volumes OCT e usa aprendizagem contrastiva supervisionada para melhorar a capacidade discriminativa dos modelos. Os experimentos realizados sugerem que a utilização de mapas de espessura da retina é mais eficaz do que a utilização dos volumes de OCT em termos de classificação e eficiência computacional. Os experimentos foram realizados utilizando a ResNet50 pré-treinada como extrator de características para classificar as imagens do conjunto de treino do dataset GAMMA, alcançando como melhores resultados uma acurácia de 0,900 e um valor Kappa de 0,855. Este trabalho utilizou apenas o conjunto de treino do dataset GAMMA.

O modelo de fusão multimodal ELF (*End-to-end Local and Global Multi-modal Fusion Framework for Glaucoma Grading*) para a classificação do glaucoma, que possui capacidade de aproveitar informações complementares entre imagens de fundo e volumes OCT, foi proposto em (LI; PUN, 2023). O ELF incorpora mecanismos de atenção local e global para explorar as informações mútuas entre as diferentes modalidades, diferenciando-se de abordagens anteriores que concatenam características multimodais.

O método foi avaliado no conjunto de treino do dataset GAMMA e reportou como melhor resultado um valor Kappa de 0,896.

Wang et al. (2023) desenvolveram a rede MSTNet, que adota uma abordagem de fusão de características combinando informações espaciais de imagens de tomografia de coerência óptica (OCT) e de retinografias. O processo de extração de características é realizado por meio da utilização da ResNet101 para as imagens de fundo de olho, enquanto uma Swin Transformer modificada é empregada para processar os volumes de OCT. A fusão de características é feita por meio de um método baseado em relações espaciais, integrando as informações extraídas de ambas as modalidades. O método foi avaliado no conjunto de treino da base GAMMA, alcançando como melhor resultado um valor Kappa de 0,892.

Wang et al. (2024) apresentam um método de aprendizagem multimodal para a detecção automatizada de doenças oculares, utilizando imagens de retinografia e imagens OCT. O estudo propõe o GeCoM-Net, que utiliza a correspondência geométrica entre as imagens de OCT e de fundo para a fusão de informações e extração de características discriminativas. Além disso, incorpora um módulo de seleção de características multi-instância (MIFS, do inglês *Multi Instance Feature Selection*) para extração eficiente de características das imagens de OCT. O modelo foi avaliado para detecção de glaucoma no dataset GAMMA, alcançando como melhor resultado um valor Kappa de 0,884 no conjunto de teste.

O modelo proposto por Kong et al. (2024) utiliza uma arquitetura de duas branches formadas pela rede DenseNet121 para extração de características de imagens de fundo de olho e OCT. Duas estratégias de fusão, L1 e MFB (*Mixed Fusion Block*), são propostas para melhorar a precisão da classificação. A fusão L1 combina características usando a normalização L1, com o objetivo de melhorar a seleção e reduzir os custos computacionais. O bloco MFB utiliza uma estratégia de fusão hierárquica para preservar atributos específicos de cada modalidade e, ao mesmo tempo, aproveitar as correlações entre características de resoluções diferentes. O método foi avaliado no conjunto de teste do dataset GAMMA e alcançou como melhor resultado um valor Kappa de 0,85.

A CRD-Net (LIU et al., 2024) é composta pelo módulo de atenção transmodal (CMA) projetado para extrair características das imagens de fundo e da fatia central do volume OCT e, ao mesmo tempo, suprimir informações irrelevantes. Esse modelo tem como objetivo comparar imagens de diferentes modalidades para tomar decisões. A arquitetura do modelo possui duas branches de extração formadas por CNNs, seguidas pelo módulo CMA, uma camada de concatenação e um classificador. O modelo foi treinado e avaliado no conjunto de treino do dataset GAMMA e alcançou como melhor resultado um valor Kappa de 0,8485.

Zou et al. (2024) propõe o modelo EyeMoSt+, que utiliza CNNs ou transformers

pré-treinados para realizar a extração de características de imagens de fundo de olho e OCTs. Na etapa seguinte, essas características são processadas por meio de *multi-evidential heads*, que combinam as predições e as incertezas geradas por cada modalidade. O processo de fusão é responsável por atualizar a média e a variância da distribuição conjunta, considerando os níveis de confiança provenientes de cada modalidade, de modo que a previsão final reflita a confiabilidade dos dados de entrada. O método foi avaliado no conjunto de treino do dataset GAMMA, alcançando como melhor resultado um valor Kappa de 0,761.

Em (WU; XUE; ZHANG, 2025) é desenvolvido um método de programação genética de múltiplas árvores (MFGP, do inglês, *Multimodal Feature Genetic programming*) para melhorar a classificação de imagens médicas multimodais. O MFGP integra estratégias de fusão de nível de característica e de nível de decisão para utilizar plenamente as características específicas da modalidade e as características multimodais. O método é avaliado no conjunto de treino da base GAMMA, superando abordagens de modalidade única. A pesquisa enfatiza a interpretabilidade da programação genética como um benefício para aplicações médicas e alcançou como melhor resultado um valor de acurácia de 0,74.

Yu et al. (2025b) desenvolveram um método que combina um algoritmo de diagnóstico por fusão de distribuição multimodal, um algoritmo de geração intermodal e uma estratégia de colaboração multitarefa. Com o objetivo de lidar com o fato de que diagnósticos unimodais negligenciam incertezas como ruído ou desalinhamento, o algoritmo de fusão proposto utiliza um codificador de distribuição de probabilidade, que transforma a predição de cada modalidade em vetores de média e variância, que são multiplicados. A variância dessa combinação é menor que a de uma única modalidade, o que diminui a incerteza do resultado. O método alcançou como melhor resultado no conjunto de treino da base GAMMA o valor de sensibilidade de 0,871.

Zhao et al. (2025) propuseram um método que realiza a extração de características a partir de imagens CFP e OCT por meio de redes neurais 2D e 3D, respectivamente. Cada modalidade é processada inicialmente por um codificador base, seguido por um codificador compartilhado e um codificador específico, com o objetivo de obter representações compartilhadas entre modalidades e específicas de cada modalidade. Para promover uma aprendizagem eficaz dessas representações e reduzir a disparidade entre modalidades, o método adota uma estratégia de regularização em múltiplos níveis. Utilizando o conjunto de treinamento da base GAMMA, o método obteve um valor máximo de Kappa de 0,865.

Em (YU et al., 2025a) foi desenvolvido um método estruturado em três níveis: um para a extração de representações a partir de retinografias, outro para OCTs e um terceiro voltado à decisão. No primeiro nível, são extraídas características globais e

pontos-chave das retinografias, que alimentam uma rede responsável pela geração de representações *tissue-aware*. Paralelamente, uma segunda rede extrai características *structure-aware*. Essas três representações são então processadas por classificadores MLPs (do inglês, *Multi Layer Perceptron*) encarregados da classificação para essa modalidade. Para as OCTs, a geração de representações inicia-se com a segmentação dos volumes utilizando o algoritmo Conditional Random Field Supervoxel, que divide os dados em supervoxels. Esses supervoxels são utilizados por uma rede para extrair características globais. As informações extraídas alimentam a rede MSAS-ViT 3D, responsável pela geração de representações *tissue-aware*. Em seguida, a rede Graph-ViT 3D seleciona os supervoxels mais relevantes para gerar representações *structure-aware*. Assim como no caso das retinografias, essas três representações são processadas por MLPs específicas para classificação da modalidade. Por fim, o nível de decisão integra as previsões geradas pelos dois ramos anteriores, consolidando o resultado final da predição. No conjunto de treinamento da base GAMMA, o método obteve como melhor resultado uma sensibilidade de 0,885.

A Tabela 1 apresenta os trabalhos relacionados que apresentaram métodos para classificação de glaucoma, apresentando inicialmente trabalhos que usaram bases da dados privadas, os que treinaram e avaliaram os métodos no conjunto de treino da base GAMMA, os que avaliaram os métodos no conjunto de teste da base GAMMA, as arquiteturas e estratégias de fusão de características, quantidade de classes da tarefa, as modalidades de imagem utilizadas para classificação e melhores resultados alcançados.

A análise dos trabalhos relacionados evidencia tendências promissoras, como a criação de bases de dados multimodais cada vez mais abrangentes, a exploração das relações entre características extraídas de diferentes modalidades de imagem e o desenvolvimento de métodos que integram predições com medidas de confiança dos modelos. Observa-se, ainda, que as estratégias de fusão seguem principalmente dois enfoques: a fusão de características (*feature-level fusion*), que combina mapas de características, e a fusão em nível de decisão (*decision-level fusion*), que integra probabilidades ou predições resultantes de diferentes classificadores. Outro aspecto relevante é a avaliação desses métodos na classificação de distintas doenças. Contudo, permanecem limitações significativas, como a baixa adoção de técnicas de explicabilidade que permitam interpretar os resultados, em especial no que diz respeito à contribuição de cada modalidade de imagem para a decisão final.

Tabela 1 – Resumo dos trabalhos relacionados.

	Trabalho	Técnica(s)	Classes	Modalidade	Resultado	Métrica
Datasets Privados	An et al. (2019)	Arquitetura multinível com a rede VGG19 como extrator de características	2	Fundo/OCT	0,96	AUC
	Mehta et al. (2021)	Ensemble de modelos DenseNet (OCT) e InceptionResnetV4 (Imagem de Fundo)	2	Fundo/OCT	0,97	
	Ma et al. (2021)	Avalia classificadores de ML que utilizam 15 dados clínicos para detecção de glaucoma	3	Fundo	0,90	
	Xiong et al. (2021)	Arquitetura multinível formada por CNNs e blocos de atenção	2	CV/OCT	0,95	
Dataset GAMMA - (Treino)	Cai et al. (2022)	Extração de mapas de espessuras de OCT e concatenação e aprendizado contrastivo.	3	Fundo/OCT	0,855	Kappa
	Li e Pun (2023)	ResNet como backbone com blocos de atenção local e global	3	Fundo/OCT	0,896	Kappa
	Wang et al. (2023)	Arquitetura com ResNet101 e Swin Transformer e fusão baseada em relações espaciais	3	Fundo/OCT	0,892	Kappa
	Liu et al. (2024)	Extração com CNNs customizadas e Módulo de atenção transmodal	3	Fundo/OCT	0,848	Kappa
	Zou et al. (2024)	Utiliza as predições e as incertezas geradas por cada modalidade	3	Fundo/OCT	0,761	Kappa
	Yu et al. (2025b)	Combina as predições de cada modelo para diminuir a incerteza do resultado.	3	Fundo/OCT	0,871	Sensibilidade
	Zhao et al. (2025)	Utiliza um codificador para obter representações compartilhadas e específicas.	3	Fundo/OCT	0,865	Kappa
	Yu et al. (2025a)	Integra representações globais, *tissue* e *structure* aware para classificação.	3	Fundo/OCT	0,885	Sensibilidade
	Wu, Xue e Zhang (2025)	Programação genética de múltiplas árvores e fusão de característica e decisão	3	Fundo/OCT	0,74	Acurácia
Dataset GAMMA - (Teste)	Fang et al. (2021)	Arquitetura Multinível com ResNet34 como *backbone* e Concatenação de características	3	Fundo/OCT	0.860	Kappa
	Li et al. (2022)	Arquitetura Multinível com ResNets como *backbone* e Concatenação Hierárquica	3	Fundo/OCT	0,875	
	Wang et al. (2024)	Extração de características das imagens com correspondência geométrica	3	Fundo/OCT	0,884	
	Kong et al. (2024)	DenseNet121 para extração e duas estratégias de fusão, L1 e MFB	3	Fundo/OCT	0,85	

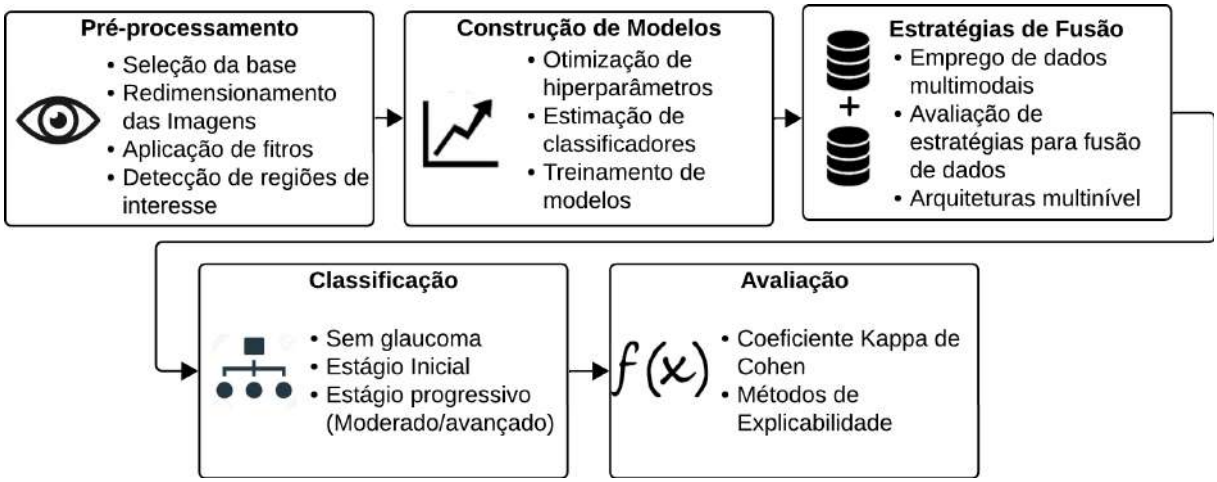
4 Método Proposto

Esta pesquisa propõe a avaliação de modelos multimodais para a fusão de características em imagens oftalmológicas, visando à melhoria do reconhecimento do padrão. Os modelos serão avaliados com imagens para o diagnóstico de glaucoma.

Este capítulo descreve o método proposto para o reconhecimento de patologias da visão com base em dados multimodais. Cabe informar que se pretende desenvolver e avaliar este método multimodal para a classificação dos estágios do glaucoma.

As etapas do estudo proposto estão apresentadas na Figura 22. A base de imagens utilizada para validar o método proposto de classificação do estágio de glaucoma é descrita na Seção 4.1. Inicialmente, foi realizada a etapa de pré-processamento na qual as imagens de fundo e os volumes OCT foram redimensionados. Foi aplicado um filtro bilateral para a eliminação dos ruídos presentes nas fatias dos volumes OCT. Em seguida, o disco óptico presente nas imagens de fundo de olho foi segmentado, resultando em uma nova imagem. Na próxima etapa, foi realizado o treinamento de modelos de classificação utilizando-se apenas uma das modalidades, de modo a obter os melhores hiperparâmetros. Em seguida, os modelos unimodais obtidos no passo anterior foram avaliados para classificação multimodal, utilizando-se estratégias de fusão de características. Por fim, os modelos foram avaliados no conjunto de teste do dataset, utilizando-se a métrica de Kappa de Cohen.

Figura 22 – Etapas do método proposto.



Fonte: imagem elaborada pelo autor.

O método proposto utiliza redes neurais convolucionais combinadas em uma arquitetura multinível para prever o estágio do glaucoma em estágio inicial ou progressivo (glaucoma intermediário e avançado). A proposta combina imagens de fundo de olho e

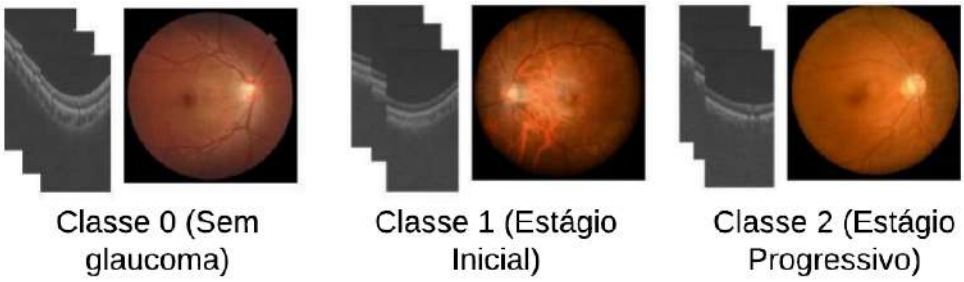
volumes de OCT, proporcionando uma análise multimodal. Neste trabalho, utilizamos arquiteturas multiníveis e estratégias de ensemble, permitindo projetar modelos que integram mais de uma modalidade de imagem médica para classificar o estágio do glaucoma.

4.1 Conjunto de dados e pré-processamento

O conjunto de dados utilizado neste trabalho foi disponibilizado aos participantes do desafio GAMMA (*Glaucoma Grading from Multi-Modality Images*) (WU et al., 2022). O conjunto de dados multimodal compreende dois exames de modalidades de imagem, imagens de fundo de olho e volumes de OCT, utilizados para o diagnóstico de glaucoma.

O conjunto de dados é composto por pares de imagens. Cada par consiste em duas modalidades de imagem, uma retinografia com resolução de (1956 x 1934 ou 2992 x 2000) e uma varredura de volume OCT com 256 fatias com resolução de 512 x 992. No total, foram disponibilizados 200 pares, sendo 100 que formam o conjunto de treinamento e 100 que formam o conjunto de teste. Ambos os conjuntos possuem pares que pertencem a uma das três classes possíveis: sem glaucoma (classe 0 - 50 pares), glaucoma em estágio inicial (classe 1 - 26 pares) e glaucoma em estágio intermediário ou avançado (classe 2 - 24 pares). A Figura 23 apresenta exemplos de pares de imagens, cada um com três cortes de OCT e uma retinografia, pertencentes a uma das três classes.

Figura 23 – Amostras do dataset GAMMA.



Fonte: (WU et al., 2022).

Foram realizados alguns pré-processamentos no conjunto de dados. As retinografias foram redimensionadas para as resoluções 128x128 e 224x224, e cada fatia dos volumes da OCT foi redimensionada para 128x128. Após o redimensionamento, foram aplicados filtros para reduzir o ruído presente nas imagens OCT. A profundidade do volume OCT foi reduzida para 64 por meio de interpolação, a fim de minimizar o custo computacional.

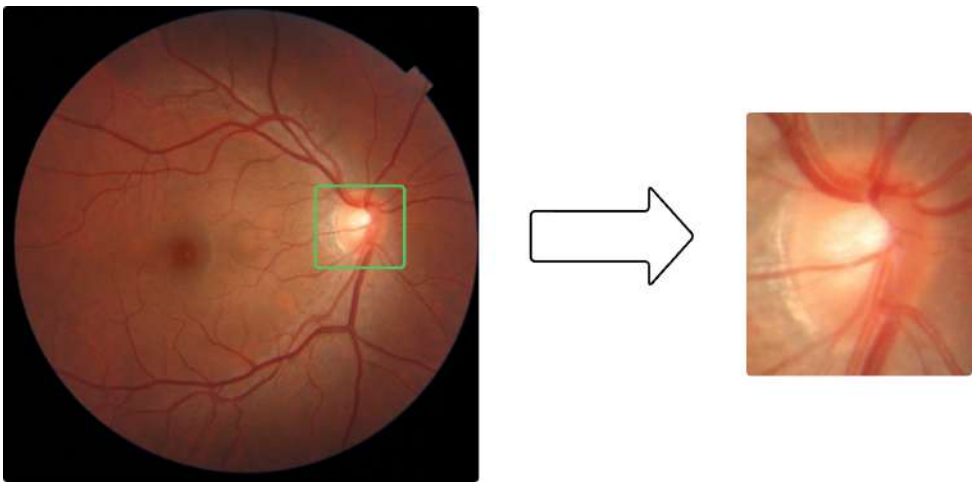
4.1.1 Captura de regiões de interesse em retinografias

Como o disco óptico é uma das regiões oculares mais relevantes para a detecção de glaucoma (WU et al., 2022), foi capturada essa região, criando uma terceira imagem exclusiva do disco óptico. Esta região de interesse foi utilizada para aumentar a capacidade de classificação dos modelos, dada a sua relevância para o diagnóstico.

Para capturar a região, foi realizada a segmentação com a rede U-NET (RONNEBERGER; FISCHER; BROX, 2015), pré-treinada em outra base de dados de retinografia, a RIMONE (BATISTA et al., 2020). Durante essa etapa, foi conduzido um processo de otimização para identificar o backbone de segmentação mais adequado entre diferentes arquiteturas baseadas em CNNs, incluindo ResNet, VGG19, DenseNet e EfficientNet. Os modelos foram comparados com base nos valores de F1-score obtidos no treinamento, sendo selecionada como melhor combinação a rede U-Net com EfficientNet como backbone.

Com base nos resultados da segmentação, a região do disco foi isolada e novas imagens foram geradas. A Figura 24 apresenta um exemplo de região de interesse contendo o nervo óptico extraído de uma retinografia.

Figura 24 – Região do nervo óptico.



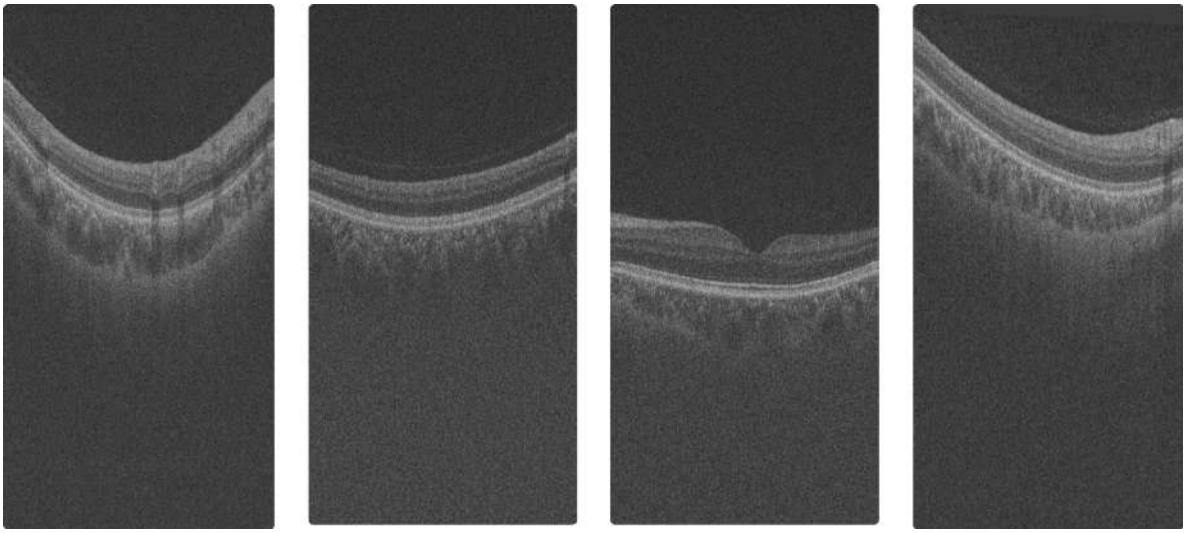
Fonte: imagem elaborada pelo autor.

4.1.2 Captura de regiões de interesse em OCTs

Com base em estudos prévios (AN et al., 2019; MWANZA et al., 2011; WU et al., 2022; GOEBEL; KRETZCHMAR-GROSS, 2002), que apontam a atrofia de camadas específicas da retina como um dos principais indicativos do glaucoma, optou-se por extrair, das imagens de OCT, a região que contém essas camadas — em especial, a camada de fibras nervosas da retina (RNFL) e a camada de células ganglionares (GCPIL). Além disso, essa escolha é reforçada pelo fato de que grande parte da imagem de OCT é

composta por *background* com alto nível de ruído, como ilustrado na Figura 25, o que pode prejudicar o desempenho dos modelos de análise.

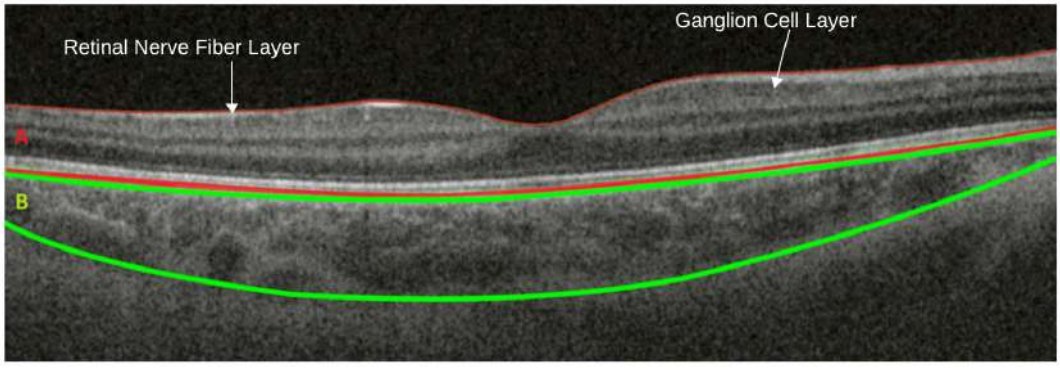
Figura 25 – Amostras de imagens OCT da base de dados utilizada nessa pesquisa.



Fonte: imagem elaborada pelo autor.

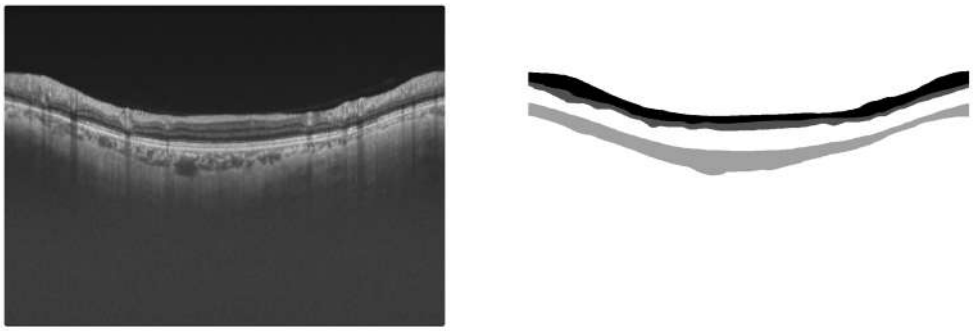
As camadas afetadas por glaucoma localizam-se na região superior das imagens de OCT, como pode ser observado na Figura 26. Deste modo, uma abordagem que utilize a região que engloba camadas retinianas acima da camada coroide pode elevar a capacidade de detecção de modelos treinados com imagens OCT. Para realizar a captura desta região de interesse, modelos de segmentação foram treinados no dataset GOALS (FANG et al., 2022), que possui máscaras das camadas RNFL, GCPIL e do coroide (Figura 27). Foi realizado um processo de otimização em busca da melhor combinação entre modelo e *backbone*, de forma semelhante ao realizado para detecção da região do disco óptico presente em retinografias. Após o processo, a combinação com os maiores valores de f1-score foi a rede U-Net, utilizando como *backbone* a rede EfficientNet-b3.

Figura 26 – Espessura total da retina (A) e espessura total do coroide (B).



Fonte: (FERNÁNDEZ-ESPINOSA et al., 2022).

Figura 27 – Amostra de imagem OCT e a máscara correspondente do dataset GOALS.



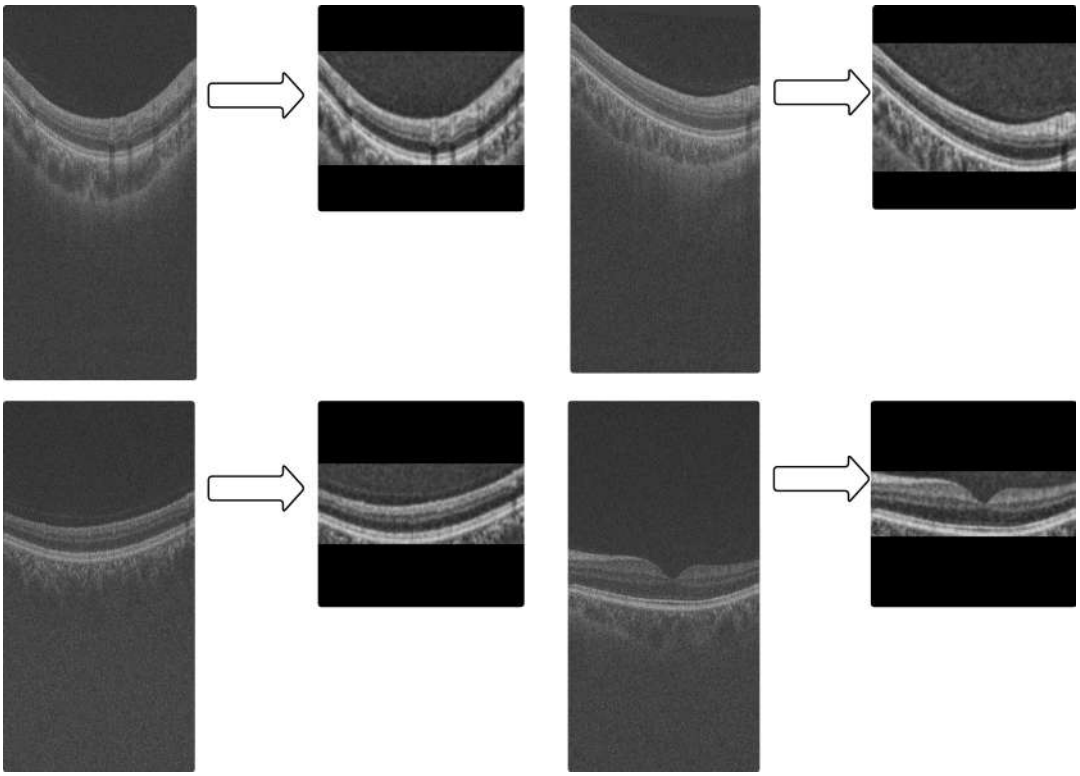
Fonte: (FANG et al., 2022).

Após o treinamento, os modelos de segmentação foram empregados para identificar e segmentar as camadas da retina localizadas acima da camada coroide. A partir das máscaras geradas, definiram-se os respectivos bounding boxes, que serviram de base para a extração das regiões de interesse (ROIs). Devido à dificuldade em segmentar adequadamente algumas amostras, foi necessário aplicar técnicas de pré-processamento — como o filtro bilateral, o CLAHE (do inglês, *Contrast Limited Adaptive Histogram Equalization*) e o desfoque gaussiano (*Gaussian Blur*) — que viabilizaram a detecção das camadas em todas as imagens dos conjuntos de treino e teste. Exemplos de regiões capturadas são apresentados na Figura 28. Como as regiões de interesse extraídas possuem dimensões variáveis dependendo da amostra, adotou-se a estratégia de centralização e preenchimento com *padding* para uniformizar as dimensões, ajustando todas as imagens para uma resolução final de 512×512 pixels. As etapas de captura de regiões de interesse de imagens OCT são apresentadas na Figura 29.

4.2 Construção dos Modelos

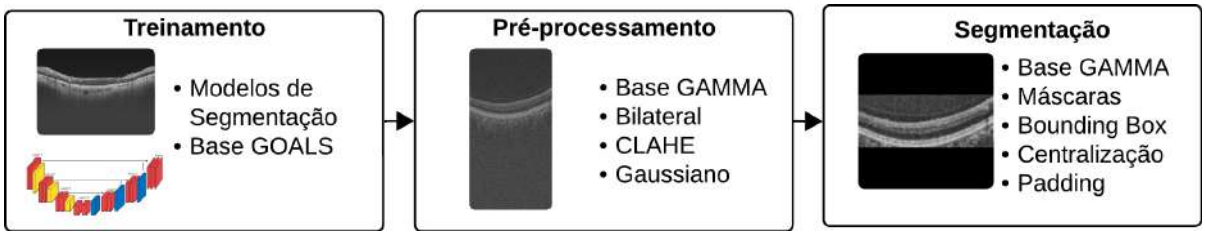
Na maioria das aplicações, as CNNs recebem imagens 2D como entrada, realizando extração de características de cada imagem para classificação. Porém, também existem CNNs 3D, que recebem volumes de entrada formados por diversas imagens denominadas fatias ou cortes. Neste trabalho, utilizamos modelos multimodais que recebem imagens de fundo de olho 2D e volumes de OCT 3D como entradas, combinando as características extraídas de cada modalidade para a classificação. Como volumes 3D foram utilizados neste trabalho, o extrator de características 3D é composto por CNNs 3D pré-treinadas, propostas por (SOLOVYEV; KALININ; GABRUSEVA, 2022). Esses modelos 3D foram obtidos a partir de CNNs 2D considerados o estado da arte. Deste modo, foi possível construir modelos com CNNs 2D e 3D, inicializados com pesos obtidos no treinamento com a base de dados Imagenet (RUSSAKOVSKY et al., 2015).

Figura 28 – Amostras de imagens OCT do conjunto de treino (à esquerda) e do conjunto de teste (à direita) e das regiões de interesse capturadas.



Fonte: imagem elaborada pelo autor.

Figura 29 – Etapas do processo de captura de regiões de interesse em imagens OCT.

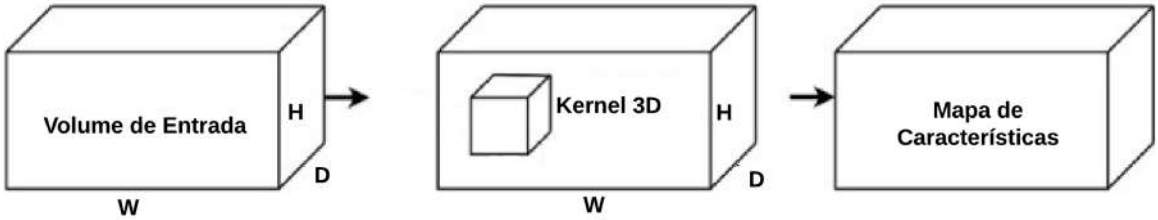


Fonte: imagem elaborada pelo autor.

A estratégia proposta por Solovyev, Kalinin e Gabruseva (2022) busca explorar a correlação entre quadros adjacentes em sequências temporais ou volumes 3D. A estratégia consiste na redistribuição uniforme dos pesos do filtro convolucional entre as fatias vizinhas. Assim, assume-se que os quadros sucessivos apresentam variações locais mínimas, o que permite compartilhar a mesma informação extraída pelo kernel 2D original de forma equilibrada ao longo do eixo temporal ou volumétrico. Para manter a escala de resposta do mapa de características inalterada, cada peso do filtro é dividido igualmente entre o número de fatias consideradas. Assim, no caso de três quadros, o kernel tridimensional (Figura 30) é definido pela equação 4.1:

$$W_1 = W_2 = W_3 = \frac{1}{3}W. \tag{4.1}$$

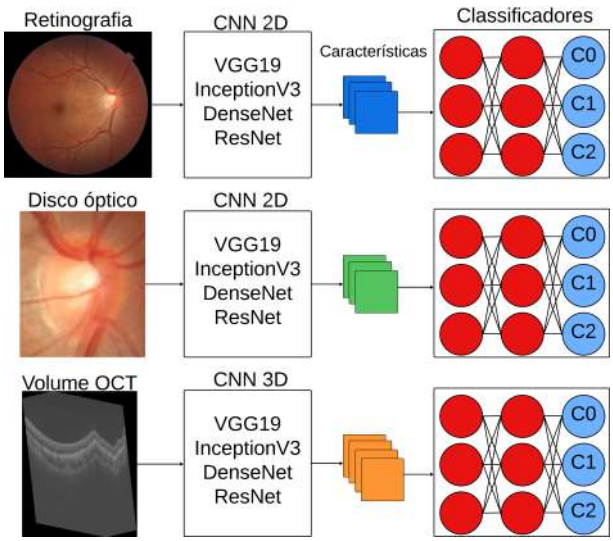
Figura 30 – Kernel tridimensional.



(Fonte: Adaptado de (SOLOVYEV; KALININ; GABRUSEVA, 2022)).

Com base nos trabalhos relacionados, avaliamos as seguintes CNNs pré-treinadas: VGG19 (SIMONYAN; ZISSERMAN, 2014), utilizada no trabalho de An et al. (2019); Inception V3 (SZEGEDY et al., 2015), utilizada no trabalho de Tian et al. (2022); DenseNet121/169 (HUANG; LIU; WEINBERGER, 2016), utilizada no trabalho de Wang et al. (2024); e ResNet50/152 (HE et al., 2016), utilizadas na maior parte dos trabalhos relacionados (FANG et al., 2021; LI et al., 2022; CAI et al., 2022; LI; PUN, 2023). Os modelos 2D receberam como entrada as retinografias e as imagens contendo a região do disco óptico, e os modelos 3D receberam como entrada os volumes de OCT. As CNNs foram importadas sem o classificador original e um novo foi adicionado, adequado à tarefa proposta. A Figura 31 apresenta os modelos 2D e 3D com as respectivas entradas.

Figura 31 – Modelos utilizados para treinamentos das retinografias, regiões do disco óptico e volumes OCTs.



Fonte: imagem elaborada pelo autor.

Na etapa de aprendizado, realizou-se a busca pelos melhores hiperparâmetros para a classificação de cada modalidade de imagem, utilizando o *framework* Optuna

(AKIBA et al., 2019a) e a estratégia de otimização Baseada em Modelo Sequencial (SMBO, do inglês *Sequential Model-Based Optimization*) (HUTTER; HOOS; LEYTON-BROWN, 2011). O SMBO difere do Grid Search e do Random Search porque considera o desempenho passado dos hiperparâmetros na busca. Em contrapartida, nos outros dois métodos, a busca é independente de avaliações anteriores. Os métodos SMBO funcionam por meio da busca do próximo conjunto de hiperparâmetros a serem avaliados na função objetivo, selecionando os hiperparâmetros que se destacam em uma função probabilística substituta, menos dispendiosa para avaliar. Caso os valores avaliados também apresentem resultados promissores na função objetivo, serão incorporados ao conjunto dos melhores hiperparâmetros.

Os seguintes hiperparâmetros foram incluídos no processo de otimização: extrator de características (CNNs), tamanho do batch, taxa de dropout, taxa de aprendizagem, número de camadas densas, número de neurônios da primeira camada densa e o divisor, utilizado para calcular o total de neurônios da segunda camada (se necessário). A Tabela 2 apresenta o espaço de busca utilizado para otimização de cada hiperparâmetro. Processos de otimização foram realizados para encontrar modelos que tenham como entrada cada modalidade de imagem: dois modelos 2D para a extração de características de retinografias e de disco óptico, e um modelo 3D para a extração de características de volumes de OCT.

Tabela 2 – Espaço de Busca.

Parâmetros	**Espaço de Busca**	**Distribuição**
CNN	VGG19 ResNet50 ResNet152 DenseNet121 DenseNet169	Categórica
Taxa de Dropout	[0,0; 0,5; step=0,1]	Uniforme Discreta
Batch size	[1; 2; 3]	Categórica
Taxa de aprendizado	[1E-5; 1E-4; 1E-3]	
Número de camadas (TC)	[1; 2]	
Número de Neurônios	[64; 128; 256; 521]	
Divisor	[2; 4; 8]	

A otimização foi realizada com base nos 100 pares rotulados que compõem o conjunto de treinamento. Dez pares foram selecionados aleatoriamente para uso como conjunto de validação. Cada modelo foi treinado por 100 épocas, utilizando como função de perda a *sparse categorical crossentropy* (Equação 4.2), adequada a tarefas de classificação multiclasse com rótulos inteiros. A estratégia de treinamento utilizada para evitar overfitting foi o *early stopping*, tendo como variáveis monitoradas a acurácia do treinamento e o loss da validação.

$$\mathcal{L}_{\text{SCCE}} = -\frac{1}{N}\sum_{i=1}^{N}\log p_{y_i}^{(i)} \tag{4.2}$$

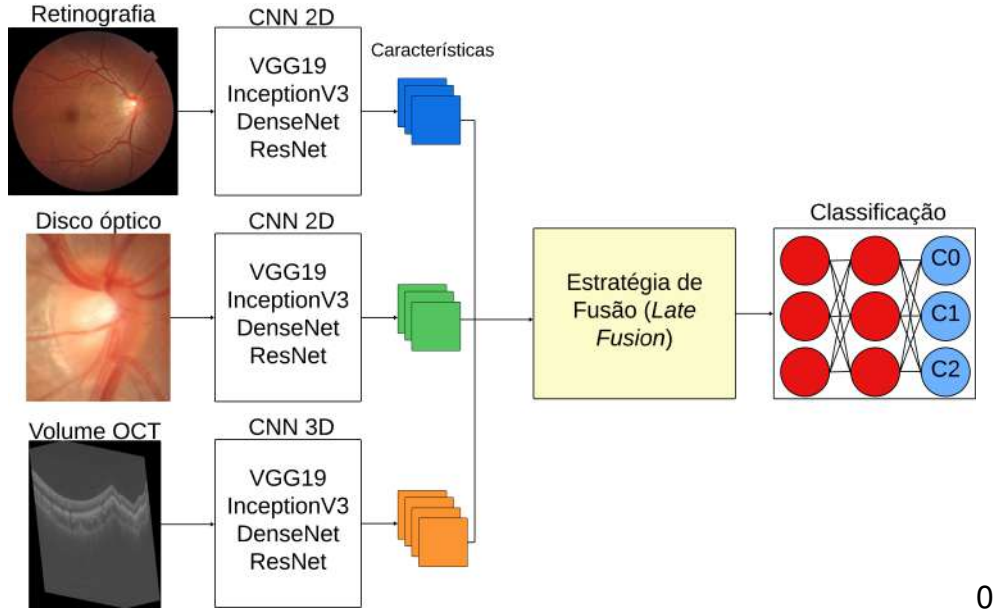
onde N é o número total de amostras no conjunto de dados, y_i é o rótulo verdadeiro da i-ésima amostra e $p_{y_i}^{(i)}$ é a probabilidade prevista pelo modelo de que a i-ésima amostra pertença à classe correta.

4.3 Fusão de Características

Para a combinação de características, foram utilizadas arquiteturas multinível, pois possibilitam a aplicação das estratégias de concatenação - *late-fusion* e de ensemble - *decision level fusion*. Não foram realizados experimentos utilizando combinação em nível de entrada (*input-level fusion*) devido às modalidades serem muito diferentes entre si.

A Figura 32 apresenta a arquitetura utilizada para aplicação da estratégia de fusão de características. As modalidades são utilizadas como entrada em cada nível das CNNs avaliadas como extratores de características. As características são utilizadas como entrada para um classificador, formado por camadas totalmente conectadas (TC), que prediz as imagens como normais (sem glaucoma) ou com glaucoma em estágio inicial ou progressivo (moderado ou avançado).

Figura 32 – Arquitetura multinível utilizada para a aplicação da concatenação de características (*late fusion*).



0

Fonte: imagem elaborada pelo autor.

A arquitetura compreende 3 níveis, o primeiro tendo como entrada retinografias, o segundo recebendo a região do nervo óptico e o último recebendo os volumes da OCT. Utilizamos o nervo óptico como segundo nível por se tratar de uma região de interesse,

pois é possível avaliar a escavação do nervo óptico, biomarcador da existência de glaucoma.

As características extraídas de cada nível são usadas para classificação. Foram utilizados modelos otimizados obtidos anteriormente, com remoção do classificador, combinados em uma arquitetura multinível. Com esta arquitetura foram avaliadas duas estratégias de fusão de características, concatenação e adição.

Na primeira estratégia, a fusão de características foi feita concatenando um vetor de características após o outro para obter um novo mapa de características, conforme usado em trabalhos anteriores (NGUYEN et al., 2019; XIONG et al., 2021).

Na segunda estratégia, os vetores de características foram combinados, adicionando-se todos os vetores para obter um novo mapa de características, de forma semelhante a redes neurais com conexões residuais. Por fim, adicionamos três camadas, duas totalmente conectadas, que utilizam a função de ativação Relu, e a última, que utiliza regressão softmax para prever a classe em normal (C0), glaucoma inicial (C1) ou glaucoma progressivo (C2).

Em seguida, foi realizado o treinamento dos modelos, avaliando-se estratégias como *warm-up learning*, em que a taxa de aprendizado tem um valor inicial baixo e aumenta gradualmente a cada época de treinamento até atingir o valor máximo definido. Não foram realizados testes com modelos de dois níveis que recebem como entrada as imagens do disco óptico e do volume OCT, pois sem retinografias completas, não haveria extração de características dos vasos sanguíneos da retina, importantes para detecção de glaucoma.

Como as imagens que compõem o conjunto de dados apresentam características visuais distintas, optou-se por adotar estratégias de ensemble (Figura 33). Neste caso, é realizada uma combinação dos resultados obtidos pelos três melhores modelos na etapa de otimização para cada imagem de entrada. Foram avaliadas as seguintes estratégias para obtenção do resultado final: ensemble da média, ensemble da moda (votação da maioria do modelos), ensemble baseado em entropia, e *stacking*.

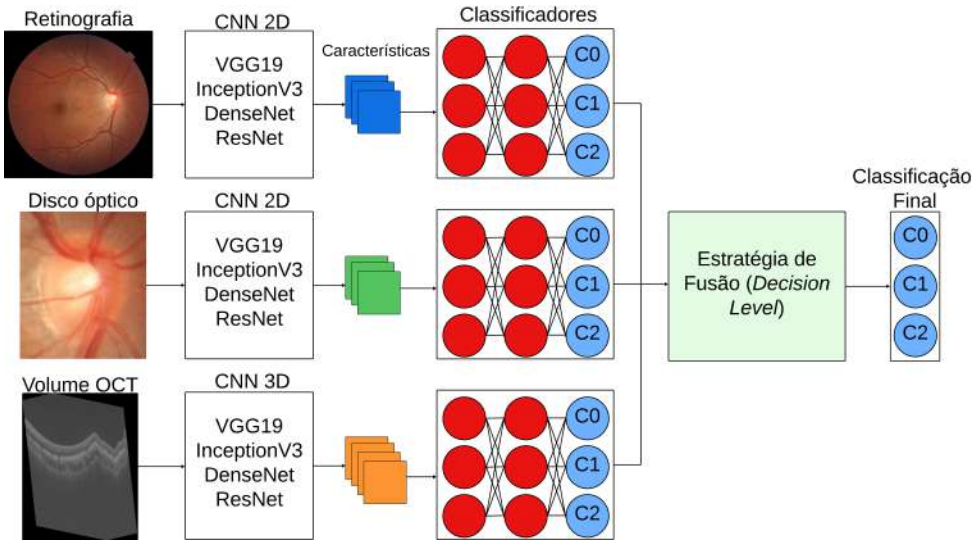
No conjunto de dados médio, considere uma tarefa com N classes e M classificadores. Sendo zij o valor do $j-\acute{e}simo$ modelo (j=1, ... M) do ith nó da última camada (i=1, ...,N). O valor médio de todos os modelos no i-ésimo nó é dado pela Equação 4.3 (NGUYEN et al., 2019).

$$v(i) = \frac{1}{M}\sum_{j=1}^{M} zji \tag{4.3}$$

onde M é o número de classificadores utilizados.

Na segunda estratégia, cada classificador prediz (vota) em uma classe, e a

Figura 33 – Arquitetura multinível utilizada para aplicação de ensemble (*decision level fusion*).



Fonte: imagem elaborada pelo autor.

decisão final é tomada pela maioria dos votos. Em termos estatísticos, o rótulo-alvo do ensemble é a moda da distribuição dos rótulos previstos individualmente (SAGI; ROKACH, 2018). Como há três classes possíveis e três classificadores, podem ocorrer empates na votação. Nesses casos, pode-se adotar o critério de maior probabilidade, selecionando-se o rótulo do classificador que apresentou a maior probabilidade na predição (Equação 4.4).

$$\text{Maior Probabilidade} = \max_{c \in \{0,1,2\}} p(c \mid x) \tag{4.4}$$

Na terceira estratégia, denominada ensemble baseado em entropia, cada classificador retorna uma distribuição de probabilidades $p_i(c \mid x)$ sobre as classes c. Calcula-se a entropia de Shannon de cada distribuição (Equação 4.5) e a decisão final é tomada com base no classificador de menor entropia (maior certeza).

Por fim, no *stacking ensemble*, as saídas probabilísticas dos modelos base $\{p_1(c \mid x), p_2(c \mid x), \ldots, p_M(c \mid x)\}$ são concatenadas e fornecidas a um meta-classificador $f_{\text{meta}}(\cdot)$, que aprende a combinar essas previsões, Equação 4.6.

$$H_i(x) = -\sum_{c=0}^{C-1} p_i(c \mid x) \log p_i(c \mid x) \tag{4.5}$$

$$\hat{y} = f_{\text{meta}}\left(p_1(c \mid x), p_2(c \mid x), \ldots, p_M(c \mid x)\right) \tag{4.6}$$

onde f_{meta} pode ser qualquer algoritmo de aprendizado supervisionado, como regressão logística, SVM ou rede neural rasa. Essa abordagem permite que o meta-classificador

aprenda padrões de erro dos modelos base e, potencialmente, supere o desempenho de qualquer modelo individual.

4.3.1 Avaliação dos Modelos

Após as etapas de treinamento, os modelos foram salvos e avaliados no conjunto de teste, composto por 100 pares não rotulados. Os resultados foram salvos em um arquivo CSV e enviados para avaliação on-line[1], que retornou o coeficiente Kappa de Cohen correspondente (Equação 4.7). Este coeficiente quantifica a concordância entre classificações, ajustando a taxa de acertos observados à probabilidade de concordância ao acaso. O coeficiente Kappa é indicado em cenários em que as classes são desbalanceadas ou quando se deseja uma medida de desempenho confiável, pois considera o acordo real em relação ao esperado pelo acaso. De acordo com (MCHUGH, 2012), o valor do Kappa varia de -1 a 1, sendo que valores próximos de 1 indicam concordância quase perfeita, valores em torno de 0 representam concordância equivalente ao acaso, e valores negativos indicam discordância sistemática. A interpretação do coeficiente de Kappa é apresentada na Tabela 3.

$$k = \frac{p_0 - p_e}{1 - p_e} \tag{4.7}$$

em que p_0 é a acurácia e p_e é a soma dos produtos dos números reais e previstos correspondentes a cada classe, dividida pelo quadrado do número total de amostras.

Tabela 3 – Interpretação do Coeficiente Kappa.

Valor	Concordância	% de dados confiáveis
0 - 0,20	Nenhuma	0 - 4%
0,21 - 0,39	Mínima	4 - 15%
0,40 - 0,59	Fraca	15 - 35%
0,60 - 0,79	Moderada	35 - 63%
0,80 - 0,90	Forte	64 - 81%
Acima de 0,90	Quase perfeita	82 - 100%

4.4 Considerações Finais

Este capítulo apresentou e descreveu em detalhes o método proposto para a classificação de glaucoma com base em dados multimodais. Foram detalhadas cada uma das etapas que compõem o método e foram apresentadas as adaptações empregadas para a classificação.

No próximo capítulo, são apresentados os resultados obtidos em cada etapa do método proposto. Além disso, são apresentadas a base de imagens aplicada, a

[1] <https://aistudio.baidu.com/competition/detail/807/0/submit-result>

configuração experimental das redes empregadas e alguns experimentos para validar as etapas do método proposto.

5 Resultados

Este capítulo apresenta os resultados experimentais de cada etapa do método proposto. A avaliação dos resultados é composta pelas seguintes etapas: 1) configuração dos experimentos; 2) resultados com modelos unimodais; 3) resultados com modelos multimodais. Após a apresentação de resultados, os mesmos serão discutidos e analisados.

5.1 Configuração dos Experimentos

Neste trabalho, propõe-se um método para a classificação do estágio do glaucoma com base em modelos multimodais. Na etapa de construção, realizou-se a otimização dos hiperparâmetros de CNNs 2D e 3D, inicializadas com pesos pré-treinados no conjunto de dados ImageNet. Os modelos receberam como entrada imagens de fundo de olho e de disco óptico, com resoluções de 128×128 e 224×224, além de volumes de OCT com dimensões 64×128×128. Para o processo de treinamento, o conjunto de treino do dataset foi dividido aleatoriamente, reservando-se 90% das amostras para treinamento e 10% para validação. As métricas de acurácia e perda (*loss*) na validação foram utilizadas como critérios para conduzir a otimização. Em seguida, os modelos com os melhores índices foram avaliados no conjunto de teste. Os rótulos deste conjunto não foram disponibilizados, e os resultados dos experimentos foram enviados à organizadora do desafio por meio da plataforma AI Studio, que retorna o valor do coeficiente Kappa de Cohen.

Na fase de fusão de características, os modelos unimodais previamente treinados foram combinados em uma arquitetura multinível, na qual as representações extraídas por cada rede serviram de entrada para a etapa de classificação. Em paralelo, também foram avaliadas estratégias de ensemble baseadas em fusão em nível de decisão (decision-level fusion), nas quais as previsões individuais de cada modelo foram consolidadas para compor o resultado final. Além dessas abordagens, investigou-se o desempenho de arquiteturas multiníveis cujos extratores de características (backbones) não haviam sido previamente treinados para classificar nenhuma das modalidades. É importante destacar que, em todas essas etapas — tanto no treinamento de modelos unimodais quanto no de modelos multimodais — foi conduzido um processo sistemático de otimização de hiperparâmetros por meio do framework Optuna (AKIBA et al., 2019b). Na seção seguinte, são apresentados os resultados experimentais obtidos com essas diferentes estratégias, evidenciando o impacto da fusão de modalidades e da otimização de hiperparâmetros no desempenho dos modelos.

5.2 Resultados com modelos unimodais

A primeira etapa do método proposto consistiu na construção de modelos unimodais, descrita em detalhes na Subseção 4.2. Nessa fase, foram desenvolvidos modelos ajustados para prever o estágio do glaucoma a partir de cada modalidade disponível. A Tabela 4 apresenta os melhores resultados obtidos, em que os modelos 2D — com e sem módulos de atenção — foram treinados para classificar retinografias, imagens da região do disco óptico, volumes de OCT e volumes de OCT compostos apenas por fatias da região de interesse (crop), conforme detalhado na Subseção 4.1.2.

Tabela 4 – Melhores resultados obtidos por modelos unimodais.

Modalidade	**Resolução**	**CNN**	**Kappa**
Imagem de Fundo	128x128	VGG19	0,799
Imagem de Fundo	224x224	Dense169	0,855
Imagem de Disco Óptico	128x128	VGG19	0,731
Imagem de Disco Óptico	224x224	Dense169	0,703
Imagem de Disco Óptico	224x224	Dense121 (Atenção)	0,764
Volume OCT	128x128x64	Dense121	0,783
Volume OCT (Crop)	128x128x64	Dense169	0,826

5.3 Resultados com modelos multimodais

Nesta etapa, foram avaliadas arquiteturas multinível, alimentadas com duas ou três modalidades. Os níveis das arquiteturas são formados por modelos unimodais previamente treinados. Para avaliar as estratégias de concatenação e adição de características (*late fusion*), o classificador dos modelos unimodais foi removido e uma nova etapa de otimização foi realizada para determinar os melhores parâmetros do novo classificador.

Foram combinados os seguintes modelos: VGG19 2D, DenseNet121 2D, DenseNet169 2D, DenseNet121 3D e DenseNet169 3D. Os resultados com modelos que utilizam duas modalidades são apresentados na Tabela 5, e os resultados com modelos que utilizam três modalidades são apresentados na Tabela 6. Como os modelos que utilizaram três modalidades alcançaram resultados melhores que os que receberam apenas duas, foram realizados experimentos em que foram inicializados com os pesos obtidos no desafio *ImageNet* e treinados diretamente na arquitetura multinível. Neste experimento, utilizou-se a rede DenseNet169 como extrator de características, uma vez que essa arquitetura obteve os melhores resultados na maioria dos testes realizados com apenas uma modalidade. O modelo foi inicialmente carregado com pesos pré-treinados no ImageNet, cuja resolução padrão é de 224×224 pixels. Por isso, as imagens de fundo

e do disco óptico foram redimensionadas para 224×224. A taxa de aprendizado foi fixada em 0,0001, pois valores menores induziram os modelos unimodais ao *overfitting*. O classificador empregado é composto por uma camada totalmente conectada com 128 neurônios, configuração idêntica à adotada nos melhores modelos unimodais. A taxa de *dropout* variou entre 0,1 e 0,3. Já os volumes de OCT foram ajustados para 64×128×128 voxels, resolução que permitiu o uso da DenseNet169 com tamanho de *batch* igual a 3, sem ocasionar problemas de memória na GPU.

Tabela 5 – Melhores resultados obtidos por modelos com duas modalidades.

CNN (Backbone)	**Pesos**	**Modalidades**	**Fusão**	**Kappa**
VGG19/VGG19	IN-PT	Fundo/Disco	Concatenação	0,809
VGG19/VGG19	IN-PT	Fundo/Disco	Adição	0,820
Dense169/VGG19	IN-PT	Fundo/Disco	Ensemble Média	0,848
VGG19/Dense121	IN-PT	Fundo/OCT	Concatenação	0,811
VGG19/Dense121	IN-PT	Fundo/OCT	Adição	0,812
Dense169/Dense121	IN-PT	Fundo/OCT	Concatenação	0,819
Dense169/Dense121	IN-PT	Fundo/OCT	Adição	0,862
Dense169/Dense121	IN-PT	Fundo/OCT	Ensemble Média	0,849
Dense169/Dense169	IN-PT	Fundo/OCT (Crop)	Concatenação	0,839
Dense169/Dense169	IN-PT	Fundo/OCT (Crop)	Adição	0,861
Dense169/Dense169	IN-PT	Fundo/OCT (Crop)	Ensemble Média	0,848

Legenda: IN-PT = Pré-treinado unimodalmente

Tabela 6 – Melhores resultados obtidos por modelos com três modalidades.

CNN (Backbone)	**Pesos**	**Modalidades**	**Fusão**	**Kappa**
VGG19/VGG19/Dense121	IN-PT	Fundo/Disco/OCT	Concatenação	0,836
VGG19/VGG19/Dense121	IN-PT	Fundo/Disco/OCT	Adição	0,839
Dense169/VGG19/Dense121	IN-PT	Fundo/Disco/OCT	Ens. Média	0,863
Dense169/VGG19/Dense121	IN-PT	Fundo/Disco/OCT	Ens. Moda	**0,886**
Dense169/Dense121/Dense169	IN-PT	Fundo/Disco/OCT (Crop)	Concatenação	0,855
Dense169/Dense121/Dense169	IN-PT	Fundo/Disco/OCT (Crop)	Adição	0,861
Dense169/Dense121/Dense169	IN-PT	Fundo/Disco/OCT (Crop)	Ens. Média	0,858
Dense169/Dense121/Dense169	IN-PT	Fundo/Disco/OCT (Crop)	Ens. Moda	0,863
Dense169/Dense121/Dense169	IN-PT	Fundo/Disco/OCT (Crop)	Ens. Entropia	0,846
Dense169/Dense121/Dense169	IN-PT	Fundo/Disco/OCT (Crop)	Ens. Stack	0,854
Dense169/Dense169/Dense169	IN-D	Fundo/Disco/OCT	Concatenação	0,805
Dense169/Dense169/Dense169	IN-D	Fundo/Disco/OCT (Crop)	Concatenação	**0,882**
Dense169/Dense169/Dense169	IN-D	Fundo/Disco/OCT	Adição	0,804
Dense169/Dense169/Dense169	IN-D	Fundo/Disco/OCT (Crop)	Adição	0,837
Dense169/Dense169/Dense169 (Atenção)	IN-D	Fundo/Disco/OCT	Concatenação	0,832
Dense169/Dense169/Dense169 (Atenção)	IN-D	Fundo/Disco/OCT (Crop)	Concatenação	0,856

Legenda: IN-PT = Pré-treinado unimodalmente; IN-D = Treinado diretamente no modelo multinível.

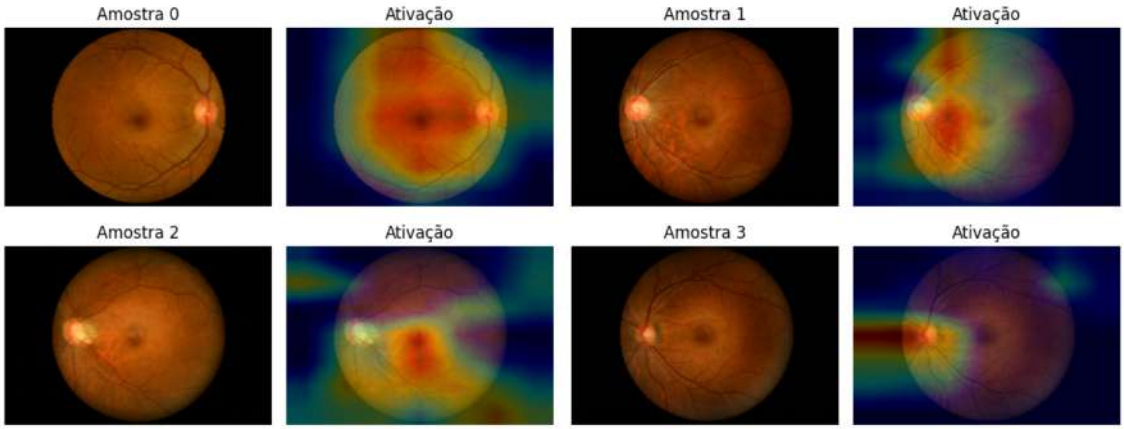
5.4 Discussão

Os resultados indicam que modelos baseados em imagens de fundo de olho apresentam desempenho superior ao dos modelos que utilizaram volumes de OCT ou imagens do disco óptico, quando considerados apenas uma modalidade. Os resultados também mostram que o uso de mais de uma modalidade eleva a capacidade de classificação do estágio de glaucoma, com os melhores resultados obtidos por modelos que utilizaram mais de uma modalidade. Entre as arquiteturas testadas, as CNNs VGG19 e DenseNet121/169 obtiveram os melhores resultados, enquanto as arquiteturas ResNet, empregadas como extratoras de características, apresentaram desempenho abaixo do esperado. Uma possível justificativa para esse resultado está relacionada à resolução das imagens: estudos prévios que usaram ResNet trabalharam com resoluções mais altas do que as adotadas neste trabalho. Para investigar mais a fundo as razões dessas diferenças, foram geradas visualizações Grad-CAM, a fim de identificar as regiões mais relevantes para a classificação.

As Figuras 34 e 35 apresentam exemplos de imagens de fundo do conjunto de teste acompanhadas de seus respectivos mapas de ativação, obtidos a partir das características exploradas pelo modelo para realizar a classificação. As amostras correspondem a retinografias com diferentes padrões no conjunto. Observa-se que, além da região do nervo óptico, o modelo também utilizou outras áreas da retina como base para a tomada de decisão. Esse comportamento pode estar relacionado ao desempenho superior dos modelos que analisam a imagem completa, uma vez que, segundo (HEMELINGS et al., 2021), redes neurais profundas podem identificar sinais de glaucoma mesmo em regiões da retina onde o nervo óptico não está presente. Adicionalmente, nota-se que áreas com maior luminosidade aparente em determinadas amostras (por exemplo, 0, 2 e 12) parecem ter atraído a atenção do modelo, o que pode justificar a relevância atribuída a essas regiões.

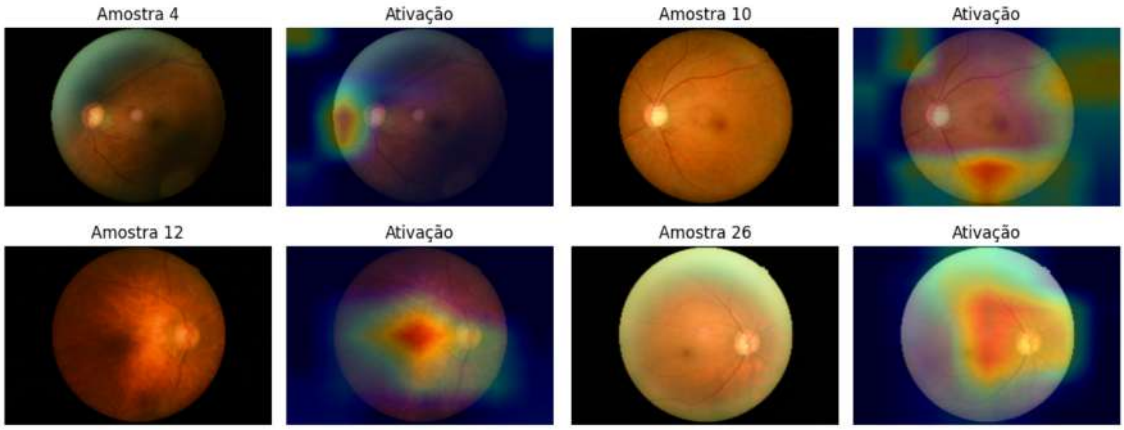
As Figuras 36 e 37 apresentam exemplos de imagens da região do nervo óptico, correspondentes às mesmas amostras de retinografia apresentadas nas Figuras 34 e 35. Os experimentos com essa modalidade resultaram em desempenho consideravelmente inferior em comparação com os modelos baseados em retinografia. Para investigar possíveis melhorias, foram avaliados mecanismos de atenção, conforme descrito na Seção 2.8. Dentre as abordagens testadas, o *Spatial Attention* proporcionou os melhores ganhos de desempenho. As figuras ilustram as ativações da última camada convolucional e como foram modificadas após a aplicação do mecanismo. Em alguns casos (amostras 0, 3, 4, 12 e 26), a atenção foi redirecionada para regiões do próprio nervo óptico, o que pode ter contribuído para o aumento de desempenho. No entanto, também foram observados efeitos adversos: na amostra 2, por exemplo, o mecanismo levou o modelo a focar em uma possível lesão próxima ao nervo óptico que não aparenta estar relacionada

Figura 34 – Amostras de imagens de fundo com os respectivos mapas de ativação.



Fonte: imagem elaborada pelo autor.

Figura 35 – Amostras de imagens de fundo com os respectivos mapas de ativação.



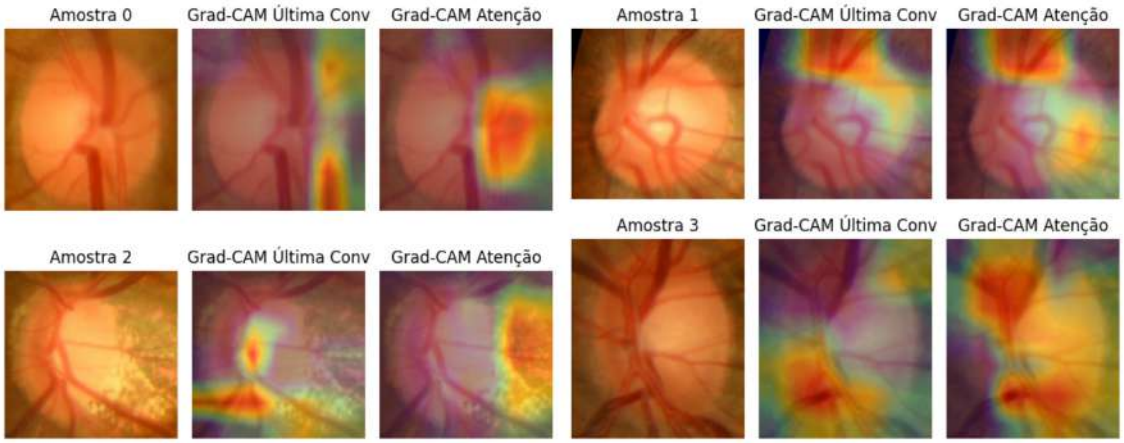
Fonte: imagem elaborada pelo autor.

ao glaucoma, o que pode introduzir erros na classificação.

Mapas de ativação também foram gerados para o conjunto de validação utilizado no treinamento, a fim de investigar se os mecanismos de atenção produzem padrões consistentes que possam ser associados às classes. A Figura 38 apresenta alguns exemplos. Nas amostras 1 e 7, observam-se lesões semelhantes às da amostra 2 da Figura 36; nesse caso, o mecanismo reduziu a atenção direcionada à lesão na camada anterior e a concentrou no nervo óptico. Em contraste, na amostra 2 ocorreu o efeito oposto. Já nas amostras 4, 5 e 9, a ativação foi direcionada ao interior do disco óptico. Na amostra 6, a alteração ocorreu principalmente nos vasos sanguíneos da região, o que levou à classificação incorreta. Essa amostra apresenta coloração da escavação semelhante à da amostra 9 (um caso de estágio inicial), o que pode ter induzido o modelo ao erro.

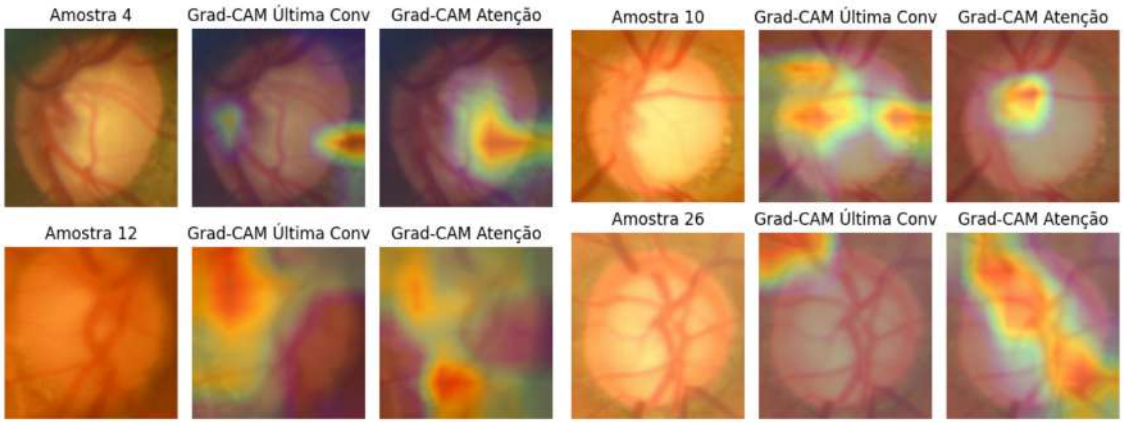
De forma análoga ao procedimento adotado para as imagens de fundo e do

Figura 36 – Amostras de imagens da região do nervo óptico com os respectivos mapas de ativação.



Fonte: imagem elaborada pelo autor.

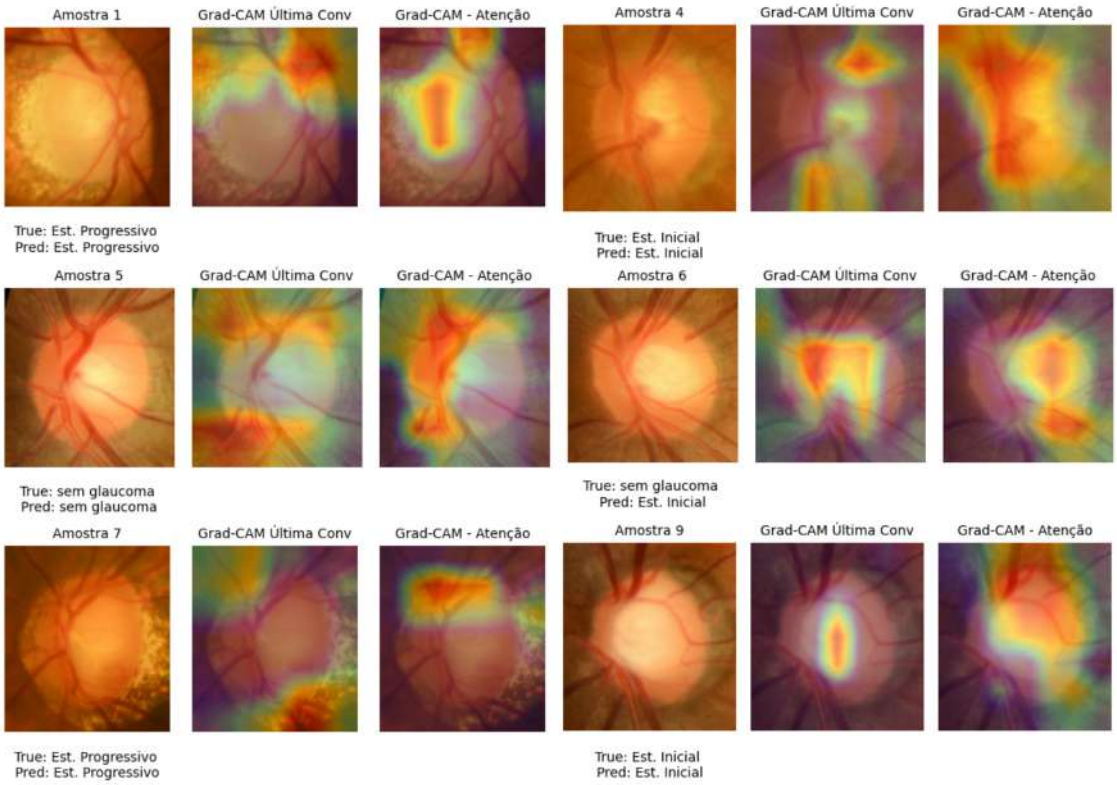
Figura 37 – Amostras de imagens da região do nervo óptico com os respectivos mapas de ativação.



Fonte: imagem elaborada pelo autor.

nervo óptico, foram gerados mapas de ativação Grad-CAM para identificar as regiões mais relevantes para a classificação dos volumes de OCT. As Figuras 39, 40 e 41 apresentam exemplos de fatias extraídas de volumes do conjunto de validação utilizado no treinamento dos modelos. Observa-se que a rede concentrou-se, predominantemente, em regiões correspondentes às camadas da retina para realizar a classificação. Alguns fatores podem ter dificultado o desempenho, como a elevada variabilidade anatômica entre as amostras do conjunto de dados e o impacto da interpolação aplicada na etapa de pré-processamento dimensional. Além disso, alterações estruturais decorrentes de outras doenças oculares podem interferir no processo de classificação, uma vez que modificam a organização das camadas retinianas. Os resultados também indicam que os modelos apresentaram uma distribuição de probabilidades mais concentrada na identificação de casos pertencentes à classe sem glaucoma, enquanto demonstraram distribuição

Figura 38 – Amostras de imagens da região do nervo óptico com os respectivos mapas de ativação, pertencentes ao conjunto de validação.



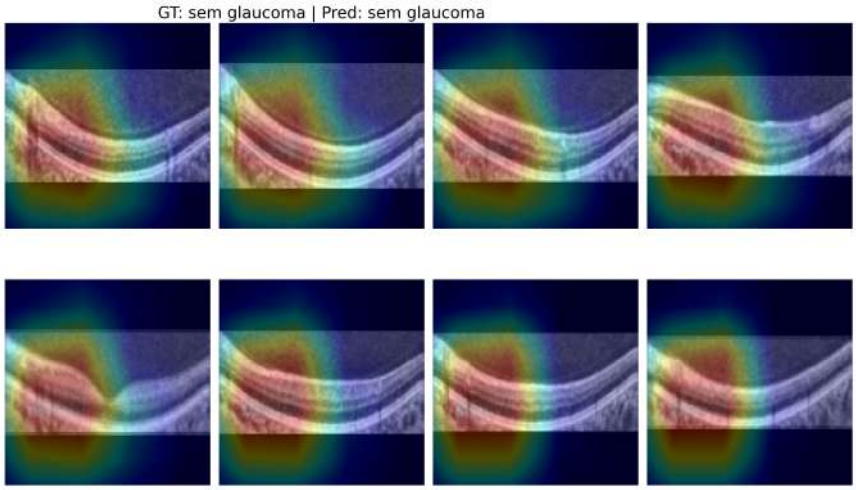
Fonte: imagem elaborada pelo autor.

mais equilibrada ao diferenciar os diferentes estágios da doença — comportamento semelhante ao observado nos experimentos com as demais modalidades de imagem.

5.5 Resultados da captura das camadas da retina em fatias de OCT

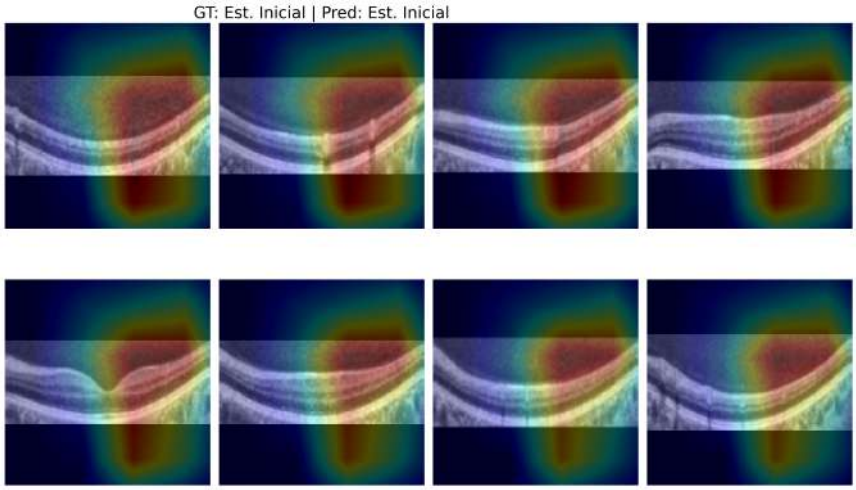
A segmentação da região das camadas da retina em fatias OCT mostrou-se uma estratégia promissora para a tarefa de classificação de glaucoma. A segmentação das camadas é uma tarefa complexa devido a casos em que doenças alteram a estrutura das camadas (LI et al., 2020). Os resultados indicam que a captura e a centralização das regiões onde as camadas retinianas são mais evidentes contribuem para elevar a capacidade discriminativa dos modelos, em comparação ao uso direto de fatias não processadas. Essa abordagem reduz a variabilidade decorrente das diferentes posições das fatias originais e concentra a análise nas estruturas mais relevantes para a detecção da doença. As figuras abaixo ilustram alguns fatores que dificultaram a captura desta região. Enquanto algumas amostras da classe ‘sem glaucoma’ não apresentam indícios de outras doenças e permitem uma segmentação clara das camadas (Figura 42), outras apresentam alterações que dificultam a identificação das estruturas, como o

Figura 39 – Exemplos de fatias OCTs com os respectivos mapas de ativação, pertencentes ao conjunto de validação (Classe Real: sem glaucoma. Previsão do modelo: sem glaucoma).



Fonte: imagem elaborada pelo autor.

Figura 40 – Exemplos de fatias OCTs com os respectivos mapas de ativação, pertencentes ao conjunto de validação (Classe Real: estágio inicial. Previsão do modelo: estágio inicial).

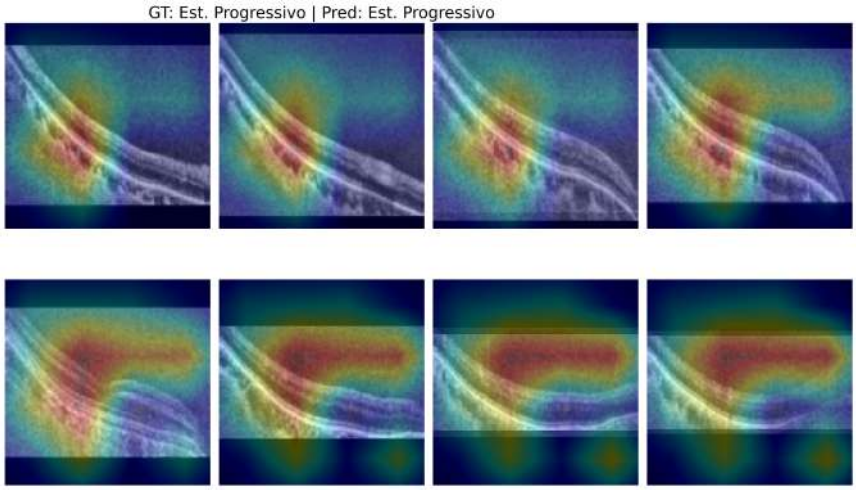


Fonte: imagem elaborada pelo autor.

comprometimento da região central próxima à fóvea (Figura 43). Neste exemplo, a região da fóvea, ou mesmo outras áreas, apresenta comprometimentos que podem estar associados a condições distintas do glaucoma, como edema ou retinopatia diabética.

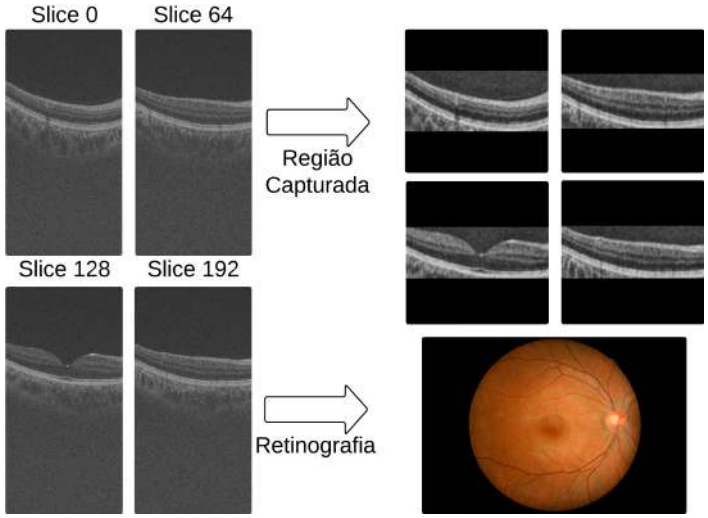
Em algumas amostras da classe ‘glaucoma moderado ou avançado’, há uma redução acentuada da espessura do epitélio pigmentar e de outras camadas retinianas (Figura 44). Além disso, neste exemplo, a qualidade da captura da imagem parece não estar boa, sendo difícil distinguir as camadas da retina do ruído presente na imagem. Na (Figura 45), pertencente à classe ’glaucoma em estágio inicial’, há sinais de edema

Figura 41 – Exemplos de fatias OCTs com os respectivos mapas de ativação, pertencentes ao conjunto de validação (Classe Real: estágio progressivo. Previsão do modelo: estágio progressivo).



Fonte: imagem elaborada pelo autor.

Figura 42 – Exemplo de captura da região das camadas da retina de uma amostra pertencente à classe sem glaucoma (Amostra do conjunto de treino).



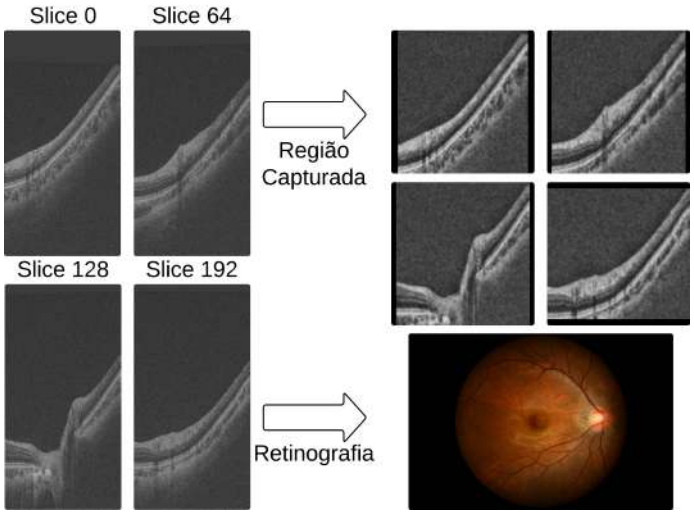
Fonte: imagem elaborada pelo autor.

macular diabético, reforçados pela lesão, que aparentemente é um exsudato.

A análise de captura das camadas da retina em OCTs foi realizada satisfatoriamente na quase totalidade das amostras de treino e de teste. No entanto, foi detectado que, em algumas amostras, o método não conseguiu capturar a região de interesse em todas as camadas do volume. As Figuras 46 e 47 apresentam esses casos.

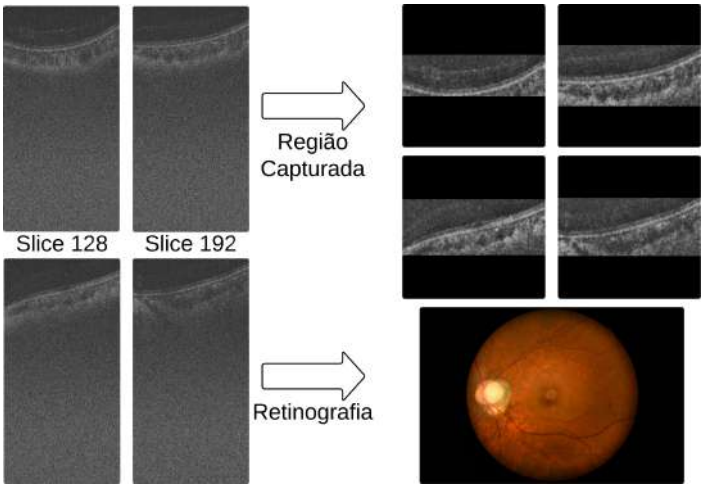
Os mapas Grad-CAM (Figura 48) da amostra ilustrada na Figura 47 foram gerados para analisar o comportamento do modelo diante de imagens com alterações nas camadas retinianas, possivelmente associadas à presença de doenças. Observa-se uma

Figura 43 – Exemplo de captura da região das camadas da retina de uma amostra pertencente à classe sem glaucoma (Amostra do conjunto de treino).



Fonte: imagem elaborada pelo autor.

Figura 44 – Exemplo de captura da região das camadas da retina de uma amostra pertencente à classe glaucoma em estágio moderado ou avançado (Amostra do conjunto de treino).



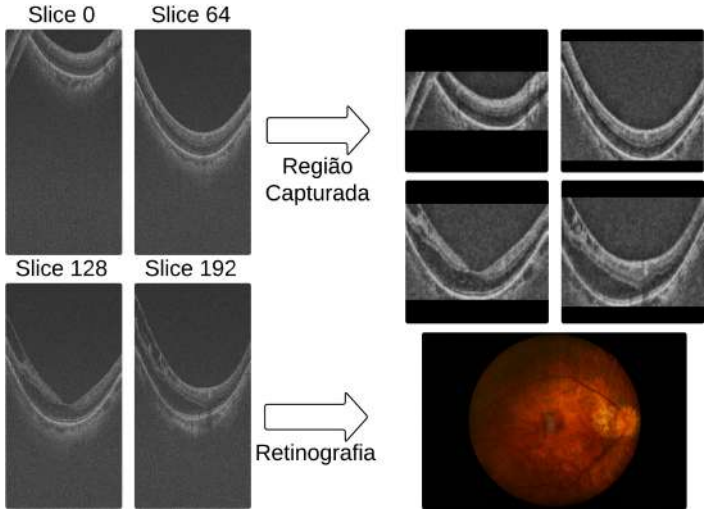
Fonte: imagem elaborada pelo autor.

expressiva variação entre as camadas retinianas ao longo das diferentes fatias, o que pode ter dificultado a capacidade de generalização do modelo.

5.6 Comparação entre as estratégias de fusão

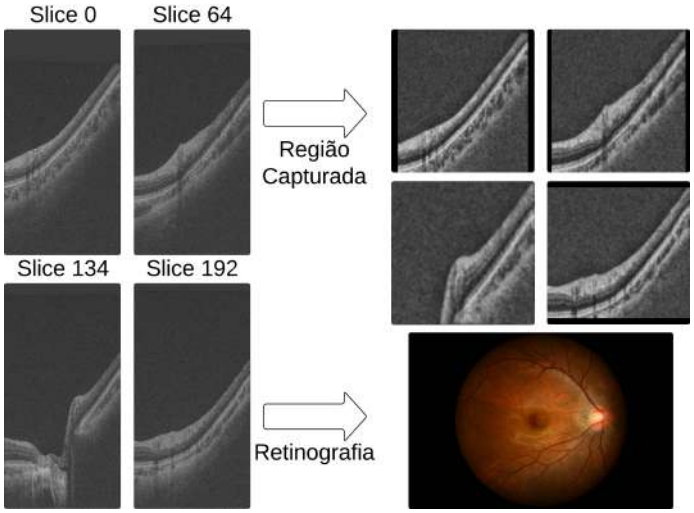
Foram avaliadas duas estratégias de combinação em modelos multimodais: a fusão de mapas de características extraídas (*late fusion*) e a fusão em nível de decisão (*decision-level fusion*). Na primeira, investigaram-se as operações de concatenação e adição das *features*, enquanto, na segunda, foram exploradas diferentes formas de

Figura 45 – Exemplo de captura da região das camadas da retina de uma amostra pertencente à classe glaucoma em estágio inicial (Amostra do conjunto de treino).



Fonte: imagem elaborada pelo autor.

Figura 46 – Exemplo de captura da região das camadas da retina de uma amostra pertencente à classe glaucoma em estágio inicial (Amostra do conjunto de treino).

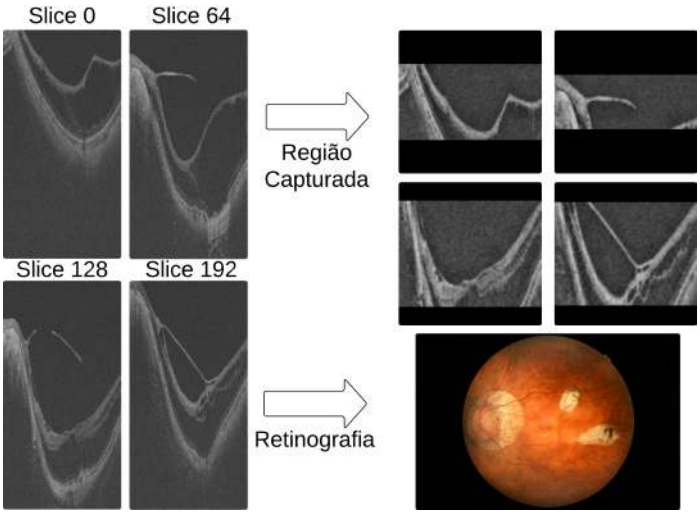


Fonte: imagem elaborada pelo autor.

combinação das distribuições de probabilidades ou das predições dos modelos, incluindo a média das probabilidades, a maioria dos votos (com desempate baseado no modelo que apresentou o maior valor de probabilidade obtido pela função softmax) e o *stacking ensemble*. Os resultados obtidos com ambas as estratégias mostraram-se semelhantes, não sendo possível determinar, de forma conclusiva, qual abordagem é mais adequada para a tarefa de classificação multimodal.

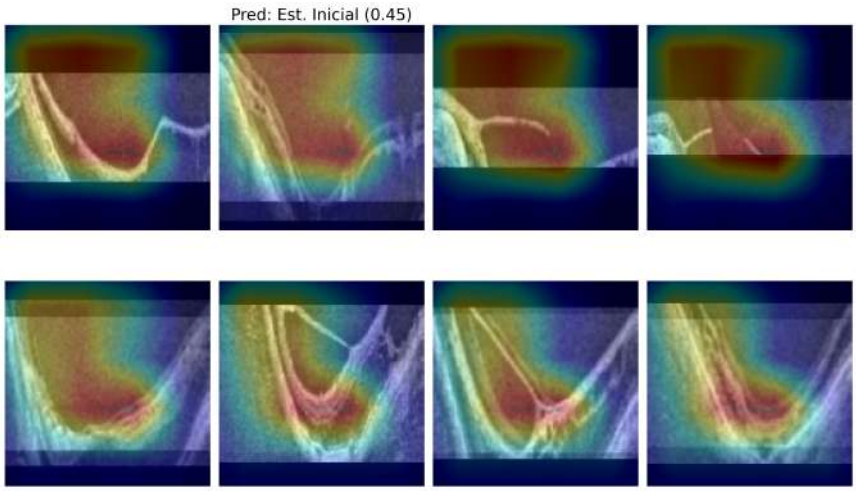
Ainda assim, algumas diferenças importantes foram observadas. A *late fusion*

Figura 47 – Exemplo de captura da região das camadas da retina de uma amostra classificada como glaucoma em estágio inicial pelo modelo ‘OCT’ e como glaucoma em estágio progressivo pelo modelo multimodal (Amostra do conjunto de teste).



Fonte: imagem elaborada pelo autor.

Figura 48 – Exemplos de fatias OCTs com os respectivos mapas de ativação, pertencentes ao conjunto de teste (Predição: Estágio Inicial).



Fonte: imagem elaborada pelo autor.

requer um número menor de parâmetros, uma vez que utiliza apenas um classificador. Além disso, a concatenação aumenta a dimensionalidade do vetor de características, o que potencialmente amplia a capacidade discriminativa do modelo. Por outro lado, a fusão em nível de decisão oferece maior flexibilidade, permitindo a exploração de diferentes esquemas de combinação para a classificação final, característica vantajosa quando se busca adaptar os modelos a diferentes cenários.

Analisando-se a distribuição de probabilidades resultante de cada modelo, é possível concluir que o modelo que utilizou imagens de fundo apresentou valores de

probabilidade mais elevados para a classe predita em relação às demais classes. Já o modelo que utiliza a região do nervo óptico apresentou valores de probabilidade menores na classificação do estágio de glaucoma, com maior desvio padrão. Como o modelo que usa imagens de fundo alcançou melhores resultados na classificação, é possível concluir que outras regiões da retina, além do nervo óptico, têm relevância considerável na detecção e classificação do estágio de glaucoma, como mostrado por (HEMELINGS et al., 2021).

De modo geral, os resultados sugerem que, embora não haja superioridade clara entre *late fusion* e *decision-level fusion*, cada abordagem apresenta vantagens complementares: enquanto a primeira favorece uma integração mais direta das informações entre modalidades, a segunda amplia a flexibilidade no processo decisório. Assim, a escolha entre as estratégias pode depender menos do desempenho e mais de critérios práticos, como custo computacional, interpretabilidade desejada e robustez diante de discordâncias entre modalidades.

5.7 Impactos de cada modalidade na classificação

Como estratégia para analisar a importância de cada modalidade na classificação, foi realizado um experimento utilizando o modelo treinado diretamente na arquitetura multinível com três níveis e concatenação como estratégia de fusão. Este modelo alcançou um valor Kappa de 0,882 no conjunto de teste (Tabela 6). O procedimento consistiu em remover uma das modalidades, substituindo as imagens correspondentes por tensores zerados de mesma dimensão. A Tabela 7 apresenta os resultados obtidos pelo modelo com as três modalidades ativas e pelas versões em que cada modalidade foi retirada individualmente.

Tabela 7 – Resultados do teste de importância de cada modalidade

Retinografia	**Disco**	**OCTs**	**Kappa**	**Total C0**	**Total C1**	**Total C2**
Presente	Presente	Presente	0,882	51	19	30
Ausente	Presente	Presente	*Overfitting*	0	0	100
Presente	Ausente	Presente	0,854	54	12	34
Presente	Presente	Ausente	0,882	51	19	30
Presente	Ausente	Ausente	0,860	54	13	33

Legenda: Total C0 = total de amostras classificadas como ‘sem glaucoma’; Total C1 = ‘estágio inicial’; Total C2 = ‘estágio progressivo’.

Os resultados do teste indicam que o melhor desempenho foi alcançado com a utilização conjunta de todas as modalidades (Kappa = 0,882). No entanto, observa-se que a combinação de retinografia e disco óptico isoladamente alcançou resultado idêntico, sugerindo que a inclusão dos volumes de OCT não acrescentou ganhos

relevantes ao processo de classificação. A análise ainda indica que a ausência da retinografia compromete significativamente a capacidade discriminativa do modelo, resultando em *overfitting* e classificação enviesada para a classe C2. Por outro lado, a retinografia utilizada isoladamente já apresentou desempenho consistente (Kappa = 0,860), reforçando sua importância como modalidade-chave. A ausência do disco óptico reduziu o Kappa para 0,854 e afetou o equilíbrio das predições, o que aponta para sua contribuição complementar no processo de decisão. Em conjunto, esses achados destacam a relevância central da retinografia, bem como o papel do disco óptico como modalidade de suporte, enquanto os volumes de OCT não se mostraram determinantes para a classificação nesta configuração experimental.

Para obter-se um cenário mais amplo da importância de cada modalidade, o mesmo teste foi realizado utilizando-se o modelo treinado diretamente na arquitetura multinível com três níveis e adição como estratégia de fusão. Este modelo alcançou o valor Kappa de 0,837 no conjunto de teste (Tabela 6). A Tabela 8 apresenta os resultados obtidos pelo modelo com as três modalidades ativas e pelas versões em que cada modalidade foi retirada individualmente.

Tabela 8 – Resultados do teste de importância de cada modalidade

Retinografia	**Disco**	**OCTs**	**Kappa**	**Total C0**	**Total C1**	**Total C2**
Presente	Presente	Presente	0,837	58	20	22
Ausente	Presente	Presente	-	83	0	17
Presente	Ausente	Presente	Overfitting	0	0	100
Presente	Presente	Ausente	0,836	54	19	27
Presente	Ausente	Ausente	Overfitting	0	0	100

Legenda: Total C0 = total de amostras classificadas como ‘sem glaucoma’; Total C1 = ‘estágio inicial’; Total C2 = ‘estágio progressivo’.

Este teste mostra a relevância da combinação das diferentes modalidades de imagem para a classificação dos estágios do glaucoma usando a adição como estratégia de fusão. O melhor desempenho, medido pelo coeficiente Kappa (0,837), foi obtido quando todas as modalidades foram utilizadas em conjunto, o que reforça a complementaridade das informações extraídas de cada fonte. De forma semelhante, a combinação de retinografia e disco óptico, mesmo na ausência dos volumes de OCT, alcançou um resultado muito próximo (Kappa = 0,836), sugerindo que essas modalidades carregam informações discriminativas suficientes para a tarefa. Em contrapartida, a exclusão de uma das modalidades (fundo ou disco) levou a desequilíbrios significativos no processo de classificação: a ausência da retinografia resultou em um viés com quase todas as amostras classificadas como C0, enquanto a ausência do disco levou ao *overfitting*, com o modelo colapsando para a classe C2. Esses resultados indicam que o uso conjunto das modalidades é importante para garantir robustez ao modelo e destacam

a importância da retinografia e do disco óptico como modalidades-chave, já que sua combinação se mostrou particularmente eficiente para a tarefa proposta.

Com base nos dois experimentos, é possível concluir que a retinografia (imagens de fundo) e o disco óptico se destacaram para a tarefa de classificação, enquanto os volumes de OCT mostraram contribuição limitada ou até instabilidade em determinadas configurações. Apesar desses indícios, os resultados ainda não permitem uma afirmação definitiva, tornando necessário ampliar a investigação com novos testes, com diferentes estratégias de fusão, a fim de validar a consistência das observações aqui apresentadas.

5.8 Impacto do conjunto de treino

Para avaliar o impacto do conjunto de treinamento no desempenho dos modelos multimodais, foi realizada uma validação cruzada com 10 folds. Nesse experimento, utilizou-se uma arquitetura multimodal baseada na CNN DenseNet169 como extrator de características, com fusão por concatenação. O classificador foi composto por uma camada densa com 128 neurônios, acompanhada de uma taxa de *dropout* de 0,3, seguida da camada final com 3 neurônios. O modelo foi treinado com taxa de aprendizado de 0,0001 e *batch size* igual a 3. Essa configuração apresentou o melhor desempenho no conjunto de teste do dataset GAMMA. Na classificação multiclasse (mais de duas classes), as métricas de precisão, sensibilidade, F1-score e AUC (*Area Under the Curve*) são calculadas individualmente para cada classe (5.1–5.4). Para obter uma medida global do desempenho do modelo, podem ser utilizadas duas estratégias: média macro, que corresponde à média aritmética simples das métricas individuais de cada classe, atribuindo o mesmo peso a todas elas; ou média ponderada (5.5), que leva em consideração o desequilíbrio do conjunto de dados, ponderando a contribuição de cada classe pelo número de instâncias verdadeiras pertencentes a ela. Os resultados obtidos em cada fold, utilizando a média ponderada, estão apresentados na Tabela 9, enquanto as matrizes de confusão correspondentes são mostradas nas Figuras 49.

$$\text{Precisão}_c = \frac{VP_c}{VP_c + FP_c} \tag{5.1}$$

$$\text{Sensibilidade}_c = \frac{VP_c}{VP_c + FN_c} \tag{5.2}$$

Onde VP é o número de verdadeiros positivos, FP é o número de falsos positivos e FN é o número de falsos negativos.

$$F1_c = \frac{2 \cdot \text{Precisão}_c \cdot \text{Recall}_c}{\text{Precisão}_c + \text{Recall}_c} \tag{5.3}$$

$$\text{AUC} = \int_0^1 \text{VPR}(\text{FPR})\, d(\text{FPR}) \tag{5.4}$$

onde:

$\text{VPR} = \frac{VP}{VP+FN}$ é a taxa de verdadeiros positivos (sensibilidade).

$\text{FPR} = \frac{FP}{FP+VN}$ é a taxa de falsos positivos.

$$\text{Métrica}_{weighted} = \frac{1}{N} \sum_{c=1}^{C} n_c \cdot \text{Métrica}_c \tag{5.5}$$

onde:

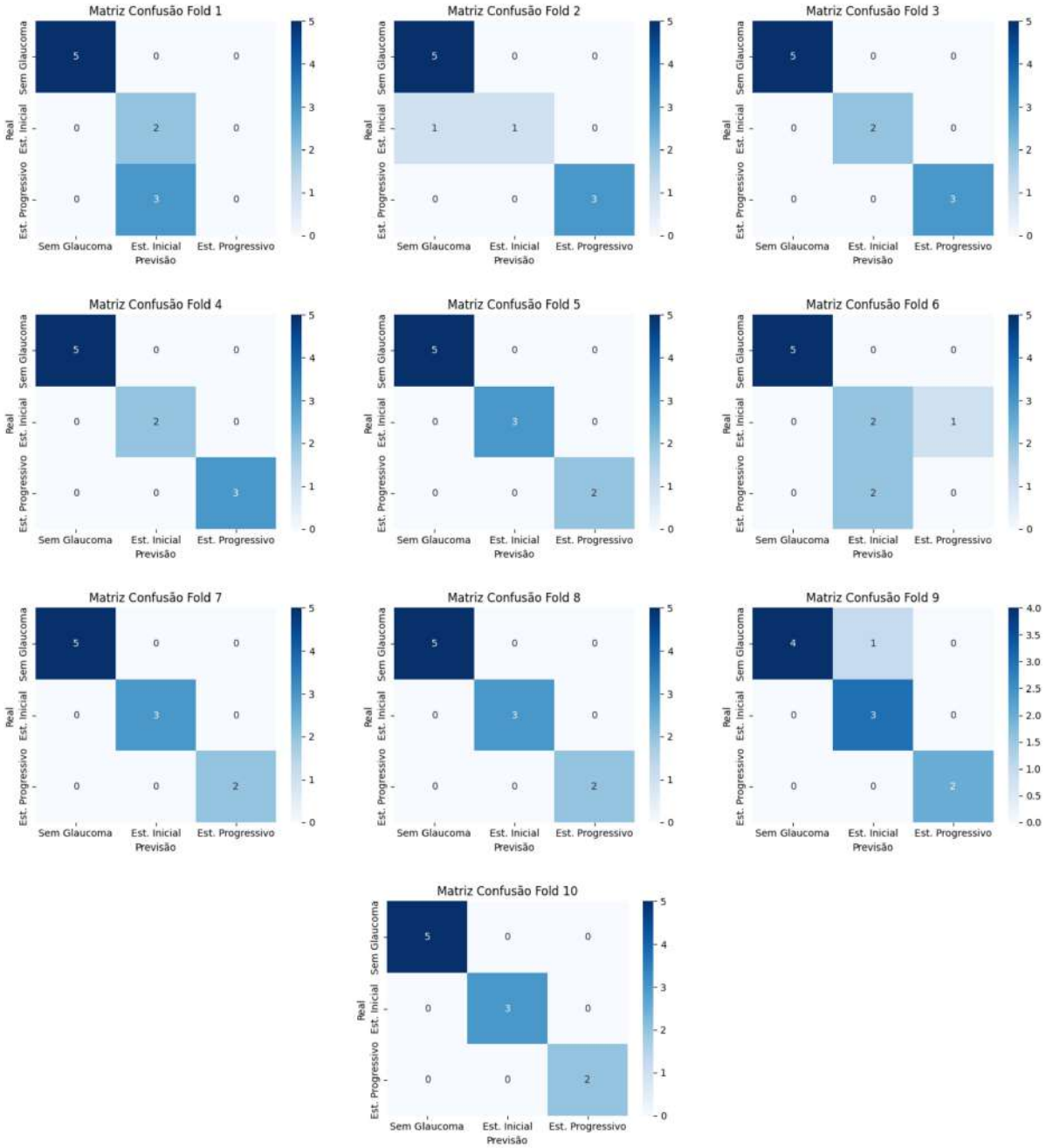
- C: número total de classes.
- n_c: quantidade de instâncias pertencentes à classe c.
- $N = \sum_{c=1}^{C} n_c$: número total de instâncias do conjunto avaliado.
- Métrica_c: valor da métrica (por exemplo, precisão, recall, F1-score ou AUC) calculada individualmente para a classe c.
- $\text{Métrica}_{weighted}$: valor da métrica ponderada, obtido pela soma das métricas de cada classe multiplicadas pelo peso correspondente $\frac{n_c}{N}$.

Tabela 9 – Resultados da validação cruzada – métricas ponderadas.

Fold	Acc	Kappa	Pre_pond	Sens_pond	F1_pond	AUC_pond
1	0.70	0.5385	0.5800	0.7000	0.6143	0.9196
2	0.90	0.8305	0.9167	0.9000	0.8879	0.9875
3	1.00	1.0000	1.0000	1.0000	1.0000	1.0000
4	1.00	1.0000	1.0000	1.0000	1.0000	1.0000
5	1.00	1.0000	1.0000	1.0000	1.0000	1.0000
6	0.70	0.5082	0.6500	0.7000	0.6714	0.9589
7	1.00	1.0000	1.0000	1.0000	1.0000	1.0000
8	1.00	1.0000	1.0000	1.0000	1.0000	1.0000
9	0.90	0.8438	0.9250	0.9000	0.9016	1.0000
10	1.00	1.0000	1.0000	1.0000	1.0000	1.0000
Média ± DP	0.92 ± 0.123	0.8721 ± 0.196	0.9072 ± 0.158	0.9200 ± 0.123	0.9075 ± 0.147	0.9866 ± 0.027

A validação cruzada também possibilitou identificar as amostras do conjunto de treino classificadas incorretamente pelo modelo. A Tabela 10 apresenta essas amostras em cada fold, destacando as classes reais e previstas. Observa-se que o modelo alcançou desempenho quase perfeito na classe sem glaucoma, cometendo apenas um erro. A Figura 50 ilustra esse caso, exibindo algumas fatias de OCT originais e processadas, juntamente com a retinografia correspondente. Nela, nota-se uma possível lesão próxima ao disco óptico, o que pode ter induzido o modelo ao erro de classificação.

Figura 49 – Matrizes relativas a cada fold da validação cruzada

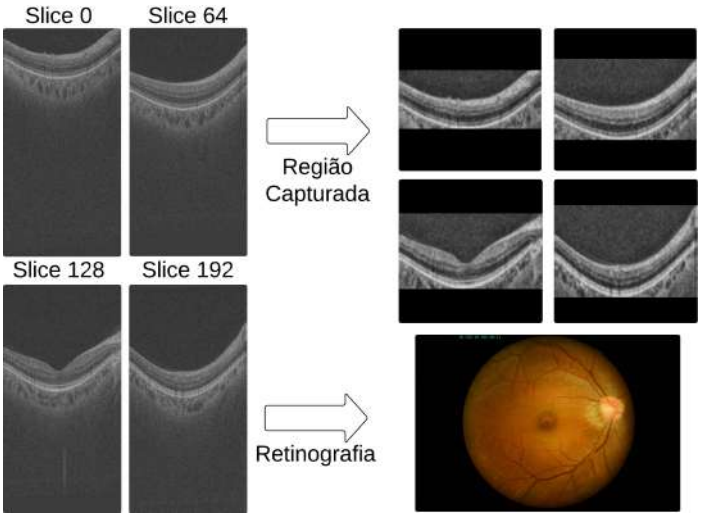


Fonte: imagem elaborada pelo autor.

Tabela 10 – Erros da validação cruzada – índice da amostra, classe real e classe prevista

Fold	Índice Amostra	Classe Real	Classe Prevista
1	0058	Estágio progressivo	Estágio Inicial
	0072	Estágio progressivo	Estágio Inicial
	0092	Estágio progressivo	Estágio Inicial
2	0009	Estágio Inicial	Sem glaucoma
6	0004	Estágio progressivo	Estágio Inicial
	0086	Estágio progressivo	Estágio Inicial
	0097	Estágio Inicial	Estágio progressivo
9	0036	Sem glaucoma	Estágio Inicial

Figura 50 – Amostra da classe ‘sem glaucoma’ classificada como ‘glaucoma em estágio inicial’.

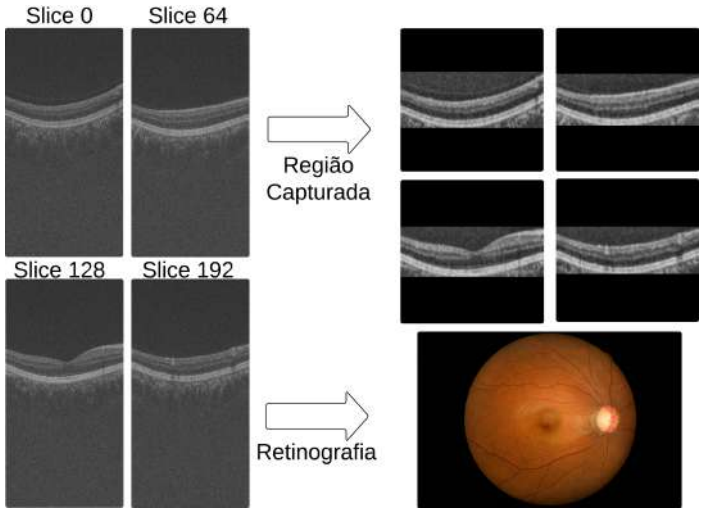


Fonte: imagem elaborada pelo autor.

Os resultados da validação indicam que a principal dificuldade do modelo está em diferenciar amostras de glaucoma em estágio inicial das em estágio progressivo (moderado ou avançado). Entre os erros de classificação, cinco ocorreram em amostras da classe estágio progressivo preditas como estágio inicial, e dois da classe estágio inicial, sendo um classificado como sem glaucoma e outro como estágio progressivo. Apesar dessas classificações incorretas, a sensibilidade do modelo pode ser considerada satisfatória, uma vez que apenas um caso de glaucoma foi incorretamente classificado como ausência da doença. Este caso é apresentado na Figura 51. Através da visualização das OCTs é possível verificar que as camadas da retina parecem não ter sofrido grandes alterações em sua estrutura ou afinamento considerável. É possível que o modelo considere que amostras com grandes alterações pertencem às classes com glaucoma. A amostra classificada como estágio progressivo, Figura 52 apresenta alterações consideráveis em sua retinografia, em relação às amostras da classe ‘sem glaucoma’.

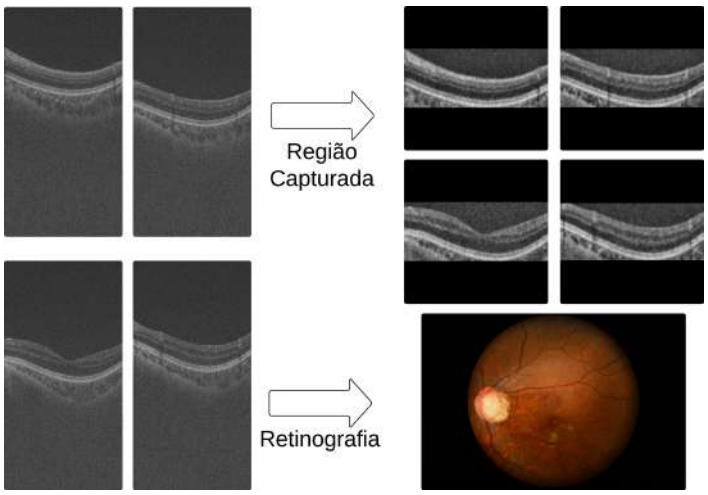
Estas alterações podem ter sido causadas pelo glaucoma, mas é possível que possam ter sido causadas por outras doenças, como a retinopatia diabética. É provável que, devido a essas alterações, a amostra tenha sido classificada como ‘estágio progressivo‘.

Figura 51 – Amostra da classe ‘estágio inicial’ classificada como ‘sem glaucoma’.



Fonte: imagem elaborada pelo autor.

Figura 52 – Amostra da classe ‘estágio inicial’ classificada como ‘estágio progressivo’.



Fonte: imagem elaborada pelo autor.

5.9 Comparação com Trabalhos relacionados

Nesta seção, são apresentadas comparações entre os resultados obtidos neste trabalho e os reportados na literatura. Para melhor organização, a análise foi dividida em duas subseções: a primeira reúne os estudos que avaliaram seus métodos utilizando o conjunto de teste, enquanto a segunda contempla aqueles que reportaram resultados avaliando o método proposto no conjunto de treinamento.

5.9.1 Comparação com trabalhos avaliados no conjunto de teste

A Tabela 11 apresenta uma comparação com os trabalhos relacionados que utilizaram o conjunto de teste do dataset para classificação de glaucoma. A métrica disponível para avaliar os métodos de classificação, o coeficiente de Kappa, é usada para medir confiabilidade entre avaliadores. De acordo com Cohen, um valor Kappa entre 0,81 e 1,00 indica um nível quase perfeito de concordância entre avaliadores. Outra interpretação proposta por McHugh (2012) sugere que valores próximos a 0,90 indicam uma concordância forte ou quase perfeita entre avaliadores. Em termos de comparação, os valores de Kappa alcançados com o método proposto estão próximos do melhor resultado alcançado no desafio GAMMA e outros trabalhos relacionados, indicando um forte nível de concordância com a classificação feita por especialistas que utilizaram outro exame, relatórios de campo visual (WU et al., 2022).

Tabela 11 – Comparação com trabalhos relacionados que avaliaram métodos no conjunto de teste do dataset GAMMA.

Trabalho	**Técnica(s)**	**Kappa**
Wang et al. (2024)	Extração de características das imagens com correspondência geométrica	0,884
Método Proposto	Arquitetura multinível baseada na DenseNet169, com foco na precisão da captura das camadas da retina em OCTs	0,882
Li et al. (2022)	Arquitetura Multinível com ResNets como *backbone* e Concatenação Hierárquica	0,875
Fang et al. (2021)	Arquitetura Multinível com ResNet34 como *backbone* e concatenação de características	0,860
Kong et al. (2024)	Arquitetura de dois níveis com DenseNet121 e fusão L1/MFB para integrar características de fundo de olho e OCT	0,850

O trabalho de Wang et al. (2024) explora a correspondência geométrica existente entre retinografias e os respectivos volumes OCT. É proposta uma arquitetura com 2 níveis de extração formados pela CNN 2D Densenet121. O método utiliza 64 fatias de cada volume OCT, no entanto, não foi explicado o critério para essa seleção inicial. Estas passam por um novo processo de seleção que é baseado na correspondência geométrica com a retinografia. Cada uma das 64 fatias tem suas características extraídas, e estas

passam por um módulo de seleção. O vetor de características selecionado é concatenado com o vetor resultante da retinografia, resultando nas características utilizadas para seleção.

Um método de combinação de características multirresolução foi apresentado em (LI et al., 2022) juntamente com uma arquitetura com 3 níveis, alcançando um valor Kappa de 0,84 como melhor resultado. As características foram extraídas por modelos ResNet pré-treinados, como (FANG et al., 2021). Apenas o modelo melhor selecionado (selecionado manualmente usando parte do conjunto de dados de treinamento) foi avaliado com o conjunto de teste. O trabalho (FANG et al., 2021) apresenta um baseline para o dataset GAMMA. Este trabalho testa modelos com as modalidades de imagem, avaliando uma CNN ResNet34 como extrator de características. A principal contribuição do trabalho foi apresentar o conjunto de dados e analisar a viabilidade de métodos de extração de características para diagnóstico.

O modelo proposto por Kong et al. (2024) emprega uma arquitetura de dois branches com DenseNet121 para analisar imagens de fundo de olho e OCT, explorando duas estratégias de fusão: L1, que melhora a seleção de características com menor custo computacional, e MFB, que preserva atributos específicos de cada modalidade e integra correlações em múltiplas resoluções.

5.9.2 Comparação com trabalhos avaliados no conjunto de treinamento

A Tabela 12 apresenta os trabalhos relacionados que avaliaram seus métodos com imagens pertencentes ao conjunto de treino (validação cruzada, *hold-out*) do dataset GAMMA, considerando ambas as modalidades de imagem (retinografia e OCT) e a tarefa de classificação em três classes (sem glaucoma, glaucoma inicial e glaucoma progressivo). Como nem todos os trabalhos reportaram a métrica Kappa, a métrica de acurácia (Acc) também foi incluída para viabilizar a comparação.

Em (YU et al., 2025a) foi desenvolvido um método estruturado em três níveis: um para retinografias, outro para OCTs e um terceiro para integração das decisões. Nas retinografias, redes distintas extraem características globais, pontos-chave e representações tissue-aware e structure-aware, que passam por MLPs de classificação. Nas OCTs, os volumes são segmentados em supervoxels por Conditional Random Field, utilizados por redes 3D (MSAS-ViT e Graph-ViT) para gerar representações globais, tissue-aware e structure-aware. Assim como nas retinografias, as representações são classificadas por MLPs próprias, e os resultados das duas modalidades são combinados em nível de decisão. O método proposto por Li e Pun (2023) apresenta uma arquitetura formada por extratores de características local e global que possuem como base a CNN Resnet50. Essas características são utilizadas por módulos de atenção global e local que realizam a concatenação 2D/3D para realizar classificação. O método é baseado na

Tabela 12 – Comparação com trabalhos relacionados que avaliaram métodos no conjunto de treino do dataset GAMMA.

Trabalho	Técnica(s)	Kappa	Acc
Yu et al. (2025a)	Integra representações globais, *tissue* e *structure* aware para classificação	-	0,940
Li e Pun (2023)	ResNet como backbone com blocos de atenção local e global	0,896	0,930
Yu et al. (2025b)	Combina as predições de cada modelo para diminuir a incerteza do resultado	-	0,930
Método Proposto	Arquitetura multinível baseada na DenseNet169, com foco na precisão da captura das camadas da retina em OCTs	0,872	0,920
Wang et al. (2023)	Arquitetura com ResNet101 e Swin Transformer e fusão baseada em relações espaciais	0,892	0,910
Liu et al. (2024)	Extração com CNNs customizadas e Módulo de atenção transmodal	0,848	0,900
Cai et al. (2022)	Extração de mapas de espessuras de OCT e concatenação e aprendizado contrastivo	0,855	0,900
Zhao et al. (2025)	Utiliza um codificador para obter representações compartilhadas e específicas.	0,865	0,860
Zou et al. (2024)	Utiliza as predições e as incertezas geradas por cada modalidade	0,761	0,860

utilização de quatro backbones para extração de características, além de utilizar imagens de fundo com resolução de 1024x1024 e volumes com fatias com resolução de 384x384, fatores que elevam o custo computacional.

Yu et al. (2025b) desenvolveram uma estratégia para lidar com incertezas inerentes aos diagnósticos unimodais, como ruído ou desalinhamento. O método combina fusão de distribuição multimodal, geração intermodal e colaboração multitarefa. Para o pré-processamento, as imagens de fundo foram redimensionadas para 512×512 e os volumes de OCT para 256×256. Cada predição é transformada em vetores de média e variância, que, ao serem multiplicados, produzem distribuições mais confiáveis, com

menor incerteza. Wang et al. (2023) propuseram a rede MSTNet, que realiza fusão de características combinando informações espaciais de retinografias e volumes de OCT. As imagens de fundo foram redimensionadas para 330×330 e as OCTs para 224×224. Para a extração, utiliza-se a ResNet101 nas imagens de fundo de olho e uma versão modificada do Swin Transformer nos volumes de OCT. As representações obtidas são então integradas por um método de fusão baseado em relações espaciais, unificando as informações das duas modalidades.

A CRD-Net (LIU et al., 2024) foi projetada com um módulo de atenção transmodal (CMA) para extrair características de retinografias e da fatia central dos volumes OCT, destacando informações relevantes e suprimindo ruídos. Sua arquitetura conta com dois níveis de extração seguidos pelo CMA, pela concatenação e por um classificador final. O método apresentado por Cai et al. (2022) é baseado na extração de mapas de espessura dos volumes de OCTs. As imagens de fundo foram redimensionadas para 1024×1024 e foram utilizados mapas de espessura com resolução de 384×384 gerados a partir dos volumes de OCT. Além disso, o modelo incorpora aprendizado contrastivo para melhorar a qualidade das representações extraídas. Embora haja um ganho significativo em termos de custo computacional, informações relevantes presentes nos volumes OCT podem não ter sido utilizadas.

Zhao et al. (2025) propuseram um método que combina redes neurais 2D e 3D para extrair características de imagens de fundo e OCTs. As imagens de fundo foram redimensionadas para 512×512 e as OCTs para 256×128. Cada modalidade passa por um codificador base, seguido por um codificador compartilhado e outro específico, de modo a capturar representações comuns e particulares. Para reduzir disparidades entre modalidades, é aplicada regularização em múltiplos níveis, o que fortalece a aprendizagem conjunta. Zou et al. (2024) apresentaram o modelo EyeMoSt+, que emprega CNNs ou transformers pré-treinados para extrair características de imagens de fundo de olho e OCTs, sendo que as primeiras foram redimensionadas para 256x256 e as segundas para 256x128. Em seguida, multi-evidential heads combinam as predições de cada modalidade, ajustando a média e a variância da distribuição conjunta conforme os níveis de confiança.

Os resultados obtidos pelo método proposto apresentam desempenho comparável aos trabalhos relacionados que reportaram as melhores performances na tarefa de classificação do estágio de glaucoma. Embora a acurácia alcançada seja inferior à observada em três dos trabalhos, considerando-se a avaliação do modelo usando o split de treino, destaca-se que a sensibilidade atingida (0,92) supera os valores reportados em (YU et al., 2025b) (0,871) e (YU et al., 2025a) (0,885), sendo que (LI; PUN, 2023) não apresentou essa métrica. Ressalta-se, entretanto, que tais trabalhos não especificam se a sensibilidade foi calculada a partir da média aritmética ou ponderada. Considerando-se

que a sensibilidade é uma métrica de elevada relevância na avaliação de modelos destinados à classificação de imagens médicas — por refletir a proporção de casos com doença que são incorretamente classificados como saudáveis, um comportamento indesejado nos modelos —, os resultados obtidos indicam potencial relevância do método proposto.

O método proposto nesta pesquisa explora duas abordagens para a classificação do estágio de glaucoma a partir de retinografias, recortes da região do nervo óptico e volumes OCT. Na primeira, foram utilizadas CNNs pré-treinadas unimodalmente, otimizadas para cada modalidade, avaliando-se distintas estratégias de combinação de características e de predições. Na segunda, aproveitando os melhores hiperparâmetros obtidos na abordagem anterior, foi implementada uma arquitetura multinível que utiliza a fusão de características de cada nível da rede, permitindo a integração das modalidades. Além disso, buscou-se compreender o processo decisório do modelo por meio da aplicação do Grad-CAM, a fim de identificar as regiões de maior relevância para as predições.

Em comparação aos métodos, a proposta difere nos seguintes aspectos. Primeiramente, as imagens utilizadas possuem baixa resolução, o que reduz o esforço computacional em relação a modelos que exploram entradas de maior dimensão. Além disso, a seleção do extrator de características em cada modalidade assegura melhor adequação da arquitetura ao problema. Outro diferencial está no pré-processamento dos volumes OCT: ao focar na região das camadas retinianas, o método evita o uso de áreas irrelevantes de background — presentes em trabalhos como (LIU et al., 2024), (ZOU et al., 2024), (ZHAO et al., 2025) e (LI; PUN, 2023) — e não realiza a captura aparentemente com critérios arbitrários (YU et al., 2025b; YU et al., 2025a; WANG et al., 2023) ou reduz o volume a uma única fatia (LIU et al., 2024). Dessa forma, são preservadas as variações anatômicas entre fatias e volumes, o que é essencial em casos de alterações patológicas.

O método apresenta algumas limitações. O uso de três níveis eleva o custo computacional, o que pode limitar sua aplicação em cenários de maior escala. Além disso, não foram explorados mecanismos de integração profunda das representações multimodais, como em (YU et al., 2025a) e (LIU et al., 2024), que buscam capturar relações mais complexas entre modalidades. Outro ponto é que a abordagem foi aplicada apenas em datasets compostos por retinografias e OCTs, diferentemente de (YU et al., 2025b) e (YU et al., 2025a), que também avaliaram OCTAs, ampliando a generalização dos modelos.

Algumas direções identificadas em trabalhos recentes podem enriquecer a linha proposta. Em (YU et al., 2025a), por exemplo, a classificação foi realizada em duas etapas — separando primeiro casos de glaucoma e não glaucoma, para em seguida diferenciar estágios da doença — estratégia que pode mitigar a dificuldade observada

neste estudo na distinção entre os estágios de glaucoma. Já (YU et al., 2025b) incorporou a geração de OCTs sintéticas a partir de retinografias, reduzindo a dependência de múltiplas modalidades, abordagem que poderia ser integrada futuramente.

Apesar dessas limitações, o modelo final apresentou desempenho competitivo. Quando avaliado no conjunto de testes da competição GAMMA — cujos rótulos não são disponibilizados publicamente e exigem a submissão dos resultados na plataforma AI Studio[1] —, o método obteve um valor de Kappa de 0,882, considerado estatisticamente como indicativo de elevado nível de concordância, o que reforça a robustez da proposta em relação aos métodos do estado da arte.

5.10 Considerações Finais

As análises realizadas permitiram verificar tanto o desempenho da abordagem quanto suas principais contribuições em relação ao estado da arte, destacando pontos fortes e aspectos que ainda podem ser aprimorados. Além disso, a interpretação dos dados possibilitou identificar limitações inerentes ao processo experimental, bem como oportunidades de investigação futura que poderão complementar e aprofundar as conclusões aqui apresentadas. Dessa forma, este capítulo cria a base empírica necessária para a conclusão do estudo, apresentada no próximo capítulo.

[1] <https://aistudio.baidu.com/competition/detail/807/0/submit-result>

6 Conclusão

Neste trabalho, foi proposto um método para a classificação dos estágios do glaucoma a partir de duas modalidades de imagem: retinografias (fundos de olho) e volumes de OCT. Para explorar o caráter multimodal dos dados, o método empregou uma arquitetura multinível e avaliou diferentes estratégias de fusão, tanto de mapas de características quanto de predições, com o objetivo de integrar de forma eficiente as informações extraídas de cada modalidade. Além disso, buscando aprimorar o desempenho dos modelos, foram utilizadas regiões específicas de interesse — o nervo óptico, nas retinografias, e as camadas retinianas, nos volumes de OCT — resultando em novas representações que ampliaram as modalidades disponíveis para a tarefa de classificação.

Com base nos experimentos realizados e nas análises apresentadas, as respostas às questões de pesquisa formuladas neste trabalho são descritas a seguir. A primeira hipótese de pesquisa trata da possibilidade de que a utilização de mais de uma modalidade de imagem médica oftalmológica possa elevar a assertividade do diagnóstico do estágio de doenças, por possibilitar aos modelos a detecção de diferentes biomarcadores e, consequentemente, predições mais precisas. Os resultados obtidos confirmam essa hipótese. Os modelos multimodais apresentaram ganhos de precisão na tarefa de classificação dos estágios do glaucoma em comparação aos modelos unimodais. Esses achados reforçam que a combinação de diferentes modalidades de imagem contribui para a melhoria da acurácia diagnóstica, uma vez que permite aos modelos capturar informações complementares relacionadas a distintos biomarcadores.

A segunda hipótese aborda se o emprego de técnicas de ensemble constitui a estratégia de fusão de características mais adequada para a tarefa de classificação utilizando um modelo multimodal, considerando que a retinografia e a tomografia de coerência óptica apresentam características visuais bastante distintas entre si. Os experimentos realizados mostraram que as estratégias de fusão baseadas em ensemble (combinação de predições) e aquelas fundamentadas na fusão de mapas de características apresentaram desempenhos semelhantes. Assim, não foi possível confirmar a superioridade das técnicas de ensemble para a tarefa de classificação. Entretanto, a análise de importância das modalidades indicou que a retinografia exerceu maior influência no processo de decisão, sugerindo que abordagens alternativas de fusão devem ser exploradas para melhor aproveitar o caráter complementar das modalidades de imagem.

A terceira questão refere-se à hipótese de que modelos de aprendizado profundo

desenvolvidos para detecção de doenças oculares em imagens de tomografia de coerência óptica possam alcançar maior desempenho diagnóstico quando treinados para focar em regiões de interesse restritas às camadas retinianas mais relevantes para cada condição específica. Os resultados confirmam essa hipótese. A utilização de volumes de OCT restritos às camadas retinianas mais relevantes proporcionou valores de Kappa superiores em relação ao uso de volumes completos. Essa evidência demonstra que a seleção de regiões de interesse pode aprimorar o desempenho diagnóstico dos modelos, representando uma direção promissora para pesquisas futuras.

De modo geral, os resultados obtidos ao longo desta pesquisa evidenciam que o uso de abordagens multimodais e o foco em regiões de interesse específicas podem contribuir significativamente para o avanço dos métodos de diagnóstico automatizado de doenças oculares. Além disso, a investigação de estratégias de fusão mais eficientes permanece como uma linha relevante para trabalhos futuros, especialmente visando ao melhor aproveitamento das informações complementares fornecidas pelas diferentes modalidades de imagem.

Os resultados obtidos trazem implicações relevantes tanto do ponto de vista metodológico quanto do clínico. Em termos de arquitetura, os modelos baseados em redes convolucionais do tipo DenseNet mostraram-se mais eficazes, reforçando seu potencial para tarefas de classificação em imagens médicas. Do ponto de vista das modalidades, os achados indicam que a retinografia exerce maior peso do que os volumes de OCT, sugerindo que estratégias que deem maior ênfase a essa modalidade podem levar a sistemas mais robustos. Embora os modelos tenham apresentado boa capacidade de distinção entre casos com e sem glaucoma, a classificação entre diferentes estágios da doença mostrou-se mais desafiadora, possivelmente devido à presença de outras condições oculares que afetam a retina e podem introduzir ruído na extração de características específicas do glaucoma. Além disso, a análise dos mapas de atenção revelou que os modelos utilizam regiões além do disco óptico nas retinografias, o que indica a possibilidade de que características adicionais da retina possam contribuir para o processo de decisão e devem ser consideradas em investigações futuras.

Este trabalho apresenta algumas limitações que devem ser consideradas na interpretação dos resultados. Primeiramente, o método foi avaliado em apenas um conjunto de dados, o que restringe a generalização dos achados e evidencia a necessidade de validação em bases mais amplas e diversas. Além disso, a adoção de redes convolucionais tridimensionais e de múltiplos níveis de extração de características elevou significativamente o custo computacional, o que pode limitar sua aplicação prática em cenários de maior escala. Outra limitação refere-se à aplicabilidade do método, que foi projetado especificamente para datasets compostos por retinografias e volumes de OCT, não sendo explorada sua adaptação a outras modalidades de imagem. Também

deve ser destacado que a dependência de duas modalidades distintas restringe a aplicabilidade do método, uma vez que, em muitos contextos clínicos, nem sempre ambas estarão disponíveis; nesse sentido, seria desejável o desenvolvimento de modelos híbridos, capazes de operar de forma flexível com uma ou mais modalidades. Por fim, não foram investigados mecanismos dedicados à integração mais explícita das relações complementares entre as modalidades, o que poderia potencialmente ampliar o desempenho alcançado.

Em síntese, este trabalho contribui para o avanço no uso de abordagens multimodais na classificação do glaucoma, explorando estratégias de fusão e o uso de regiões específicas de interesse em retinografias e OCTs. Apesar das limitações identificadas, os resultados obtidos indicam o potencial da combinação de modalidades de imagem para apoiar o diagnóstico clínico e abrem caminho para investigações futuras que ampliem a robustez e a aplicabilidade dos modelos propostos.

6.1 Contribuições

Finalmente, as principais contribuições do método proposto desenvolvido nesta tese são descritas a seguir:

1. Um modelo otimizado para classificação do estágio de severidade de glaucoma que utiliza duas modalidades de imagem médica, retinografia e OCT, e emprega estratégias de combinação de características;
2. Avaliação do impacto das modalidades de imagens de entrada para o resultado do modelo;
3. Avaliação de estratégia de fusões de características em modelos multimodais sobre imagens de OCT e retinografia;
4. Avaliação do impacto da pré-segmentação de estruturas do olho presentes nas imagens como forma de melhoria do diagnóstico.

6.2 Trabalhos Futuros

1. Realizar a validação do método em outros conjuntos de dados, de modo a verificar sua capacidade de generalização e robustez;
2. Investigar o uso de uma arquitetura em dois estágios, na qual um classificador distingue entre casos com e sem glaucoma, e outro seja responsável por determinar o estágio da doença apenas nas amostras positivas, avaliando o impacto dessa estratégia no desempenho global;

3. Explorar diferentes modelos e técnicas de aprendizado, incluindo abordagens recentes como o aprendizado contrastivo, que podem favorecer a extração de representações mais discriminativas;

4. Desenvolver e avaliar mecanismos de fusão mais profundos e explícitos, capazes de integrar de forma mais eficaz as informações complementares provenientes das modalidades de retinografia e OCT;

5. Examinar o potencial do uso de projeções bidimensionais derivadas de volumes de OCT, como a estratégia proposta em (LI et al., 2020), para a tarefa de detecção de glaucoma;

6. Investigar se o método proposto para a captura da região das camadas retinianas em OCTs pode também contribuir para a detecção de outras condições oftalmológicas, como o edema macular diabético.

6.3 Produções Científicas

A Tabela 13 apresenta o artigo publicado diretamente relacionado ao método proposto para classificação de estágio de glaucoma utilizando imagens multimodais. Além disso, a Tabela 14 lista os artigos científicos publicados e submetidos em que houve participação como autor ou co-autor em outras aplicações de processamento de imagens e visão computacional desde o início do doutorado.

Tabela 13 – Produções científicas em relação ao método proposto para classificação de doenças da retina utilizando imagens multimodais.

Artigo	**Tipo**	**Qualis**	**Status**
Glaucoma grading using multimodal imaging and multilevel CNN. Em: IEEE Latin America Transactions. Ano: 2023.	Periódico	A4	Publicado
Evaluation of the vision mamba model for detecting diabetic retinopathy. Em: Procedia Computer Science. Ano: 2024.	Periódico	A3	Publicado
Multilevel CNN for anterior chamber angle classification using AS-OCT images. Em: International Journal of Innovative Computing and Applications. Ano: 2023	Periódico	B2	Publicado

Tabela 14 – Produções científicas em outras aplicações de processamento de imagens e visão computacional.

Artigo	**Tipo**	**Qualis**	**Participação**	**Status**
Glaucoma grading using fundus images. Em: EAI International Conference on Wireless Mobile Communication and Healthcare. Ano: 2023	Conferência	B4	Orientador	Publicado
Multiple instance learning in medical imaging: A systematic review. Em: IEEE Access Ano: 2024.	Periódico	A1	Co-autor	Publicado
Attention Mechanisms in Deep Neural Networks for Diabetic Retinopathy Detection in OCT: A Systematic Review. Em: IEEE Access. Ano: 2025.	Periódico	A1	Co-autor	Submetido
Detection of Diabetic Macular Edema in Optical Coherence Tomography Using Active Learning Optimized for TPU v2-8. Em: International Conference on Health and Social Care Information Systems and Technologies. Ano: 2025.	Conferência	B4	Co-orientador	Aceito para publicação
Segmentation of Retinal Layers in OCT Images Using Deep Learning Methods. Em: International Conference on Health and Social Care Information Systems and Technologies. Ano: 2025.	Conferência	B4	Co-orientador	Aceito para publicação
Lesion Segmentation Associated with Diabetic Retinopathy Using Deep Learning Methods. Em: International Conference on Health and Social Care Information Systems and Technologies. Ano: 2025.	Conferência	B4	Co-orientador	Aceito para publicação
Automatic Optic Nerve Segmentation in Retinal Photographs for Glaucoma Detection Using Convolutional Neural Network. Em: International Conference on Health and Social Care Information Systems and Technologies. Ano: 2025.	Conferência	B4	Co-orientador	Aceito para publicação

Referências

AKIBA, T.; SANO, S.; YANASE, T.; OHTA, T.; KOYAMA, M. **Optuna: A Next-generation Hyperparameter Optimization Framework**. 2019. Disponível em: <https://arxiv.org/abs/1907.10902>.

AKIBA, T.; SANO, S.; YANASE, T.; OHTA, T.; KOYAMA, M. Optuna: A next-generation hyperparameter optimization framework. In: **Proceedings of the 25rd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining**. [S.l.: s.n.], 2019.

ALIPOUR, S. H. M.; RABBANI, H.; AKHLAGHI, M. A new combined method based on curvelet transform and morphological operators for automatic detection of foveal avascular zone. **Signal, Image and Video Processing**, Springer, v. 8, p. 205–222, 2014.

AN, G.; OMODAKA, K.; HASHIMOTO, K.; TSUDA, S.; SHIGA, Y.; TAKADA, N.; KIKAWA, T.; YOKOTA, H.; AKIBA, M.; NAKAZAWA, T. Glaucoma diagnosis with machine learning based on optical coherence tomography and color fundus images. **Journal of healthcare engineering**, Hindawi, v. 2019, 2019.

ARAÚJO, F.; CARNEIRO, A.; SILVA, R.; MEDEIROS, F.; USHIZIMA, D. Redes neurais convolucionais com tensorflow:teoria e prática. v. 1, p. 382–406, 2017.

BATISTA, F. J. F.; DIAZ-ALEMAN, T.; SIGUT, J.; ALAYON, S.; ARNAY, R.; ANGEL-PEREIRA, D. Rim-one dl: A unified retinal image database for assessing glaucoma using deep learning. **Image Analysis & Stereology**, v. 39, n. 3, p. 161–167, 2020.

BISHOP, C. M. **Pattern Recognition and Machine Learning (Information Science and Statistics)**. Berlin, Heidelberg: Springer-Verlag, 2006. ISBN 0387310738.

BOULAHIA, S. Y.; AMAMRA, A.; MADI, M. R.; DAIKH, S. Early, intermediate and late fusion strategies for robust deep learning-based multimodal action recognition. **Machine Vision and Applications**, Springer, v. 32, n. 6, p. 121, 2021.

CAI, Z.; LIN, L.; HE, H.; TANG, X. Corolla: an efficient multi-modality fusion framework with supervised contrastive learning for glaucoma grading. In: IEEE. **2022 IEEE 19th International Symposium on Biomedical Imaging (ISBI)**. [S.l.], 2022. p. 1–4.

CBO. **Visão em Foco**. 2017. Acessado em 15 de junho de 2025. Disponível em: <https://visaoemfoco.org.br/revista/14/80-dos-portadores-de-glaucoma-n%E3o-apresentam-sintomas>.

CBO. **A importância de combater e prevenir a cegueira**. 2020. Acessado em 15 de junho de 2025. Disponível em: <https://visaoemfoco.org.br/noticia/a-importancia-de-combater-e-prevenir-a-cegueira1657730359>.

CBO. Censo oftalmologico 2021. Conselho Brasileiro de Oftalmologi, 2021. <https://cbo.net.br/2020/admin/docs_upload/034327Censocbo2021.pdf>.

CHEN, T. C.; HOGUET, A.; JUNK, A. K.; NOURI-MAHDAVI, K.; RADHAKRISHNAN, S.; TAKUSAGAWA, H. L.; CHEN, P. P. Spectral-domain oct: helping the clinician diagnose

glaucoma: a report by the american academy of ophthalmology. **Ophthalmology**, Elsevier, v. 125, n. 11, p. 1817–1827, 2018.

CHUA, J.; SIM, R.; TAN, B.; WONG, D.; YAO, X.; LIU, X.; TING, D. S.; SCHMIDL, D.; ANG, M.; GARHÖFER, G. et al. Optical coherence tomography angiography in diabetes and diabetic retinopathy. **Journal of Clinical Medicine**, MDPI, v. 9, n. 6, p. 1723, 2020.

ELGAFI, M.; SHARAFELDEEN, A.; ELNAKIB, A.; ELGARAYHI, A.; ALGHAMDI, N. S.; SALLAH, M.; EL-BAZ, A. Detection of diabetic retinopathy using extracted 3d features from oct images. **Sensors**, MDPI, v. 22, n. 20, p. 7833, 2022.

FANG, H.; LI, F.; FU, H.; WU, J.; ZHANG, X.; XU, Y. Dataset and evaluation algorithm design for goals challenge. In: SPRINGER. **International Workshop on Ophthalmic Medical Image Analysis**. [S.l.], 2022. p. 135–142.

FANG, H.; SHANG, F.; FU, H.; LI, F.; ZHANG, X.; XU, Y. Multi-modality images analysis: A baseline for glaucoma grading via deep learning. In: SPRINGER. **International Workshop on Ophthalmic Medical Image Analysis**. [S.l.], 2021. p. 139–147.

FERNÁNDEZ-ESPINOSA, G.; ORDUNA-HOSPITAL, E.; BONED-MURILLO, A.; DIAZ-BARREDA, M. D.; SANCHEZ-CANO, A.; SOPEÑA-PINILLA, M.; PINILLA, I. Choroidal and retinal thicknesses in type 2 diabetes mellitus with moderate diabetic retinopathy measured by swept source oct. **Biomedicines**, MDPI, v. 10, n. 9, p. 2314, 2022.

GOEBEL, W.; KRETZCHMAR-GROSS, T. Retinal thickness in diabetic retinopathy: a study using optical coherence tomography (oct). **Retina**, LWW, v. 22, n. 6, p. 759–767, 2002.

GOODFELLOW, I.; BENGIO, Y.; COURVILLE, A. **Deep Learning**. Cambridge: MIT Press, 2016. <http://www.deeplearningbook.org>.

HE, K.; ZHANG, X.; REN, S.; SUN, J. Deep residual learning for image recognition. In: **2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)**. Los Alamitos, CA, USA: IEEE Computer Society, 2016. p. 770–778. ISSN 1063-6919. Disponível em: <https://doi.ieeecomputersociety.org/10.1109/CVPR.2016.90>.

HEMELINGS, R.; ELEN, B.; BARBOSA-BREDA, J.; BLASCHKO, M. B.; BOEVER, P. D.; STALMANS, I. Deep learning on fundus images detects glaucoma beyond the optic disc. **Scientific reports**, Nature Publishing Group UK London, v. 11, n. 1, p. 20313, 2021.

HERVELLA, A. S.; ROUCO, J.; NOVO, J.; ORTEGA, M. Retinal microaneurysms detection using adversarial pre-training with unlabeled multimodal images. **Information Fusion**, Elsevier, v. 79, p. 146–161, 2022.

HO, T. K. Random decision forests. In: IEEE. **Proceedings of 3rd international conference on document analysis and recognition**. Montreal, 1995. v. 1, p. 278–282.

HU, J.; SHEN, L.; SUN, G. Squeeze-and-excitation networks. In: **Proceedings of the IEEE conference on computer vision and pattern recognition**. [S.l.: s.n.], 2018. p. 7132–7141.

HUANG, G.; LIU, Z.; WEINBERGER, K. Q. Densely connected convolutional networks. **CoRR**, abs/1608.06993, 2016. Disponível em: <http://arxiv.org/abs/1608.06993>.

HUTTER, F.; HOOS, H. H.; LEYTON-BROWN, K. Sequential model-based optimization for general algorithm configuration. In: SPRINGER. **Learning and Intelligent Optimization: 5th International Conference, LION 5, Rome, Italy, January 17-21, 2011. Selected Papers 5**. [S.l.], 2011. p. 507–523.

KONG, Y.; ZHANG, W.; LU, S.; LI, H. A classification model for glaucoma grading using multi-modal image fusion strategies. In: IEEE. **2024 IEEE 19th Conference on Industrial Electronics and Applications (ICIEA)**. [S.l.], 2024. p. 1–6.

KULYABIN, M.; ZHDANOV, A.; NIKIFOROVA, A.; STEPICHEV, A.; KUZNETSOVA, A.; RONKIN, M.; BORISOV, V.; BOGACHEV, A.; KOROTKICH, S.; CONSTABLE, P. A. et al. Octdl: Optical coherence tomography dataset for image-based deep learning methods. **Scientific data**, Nature Publishing Group UK London, v. 11, n. 1, p. 365, 2024.

LI, M.; CHEN, Y.; JI, Z.; XIE, K.; YUAN, S.; CHEN, Q.; LI, S. Image projection network: 3d to 2d image segmentation in octa images. **IEEE Transactions on Medical Imaging**, IEEE, v. 39, n. 11, p. 3343–3354, 2020.

LI, W.; PUN, C.-M. Elf: An end-to-end local and global multimodal fusion framework for glaucoma grading. In: IEEE. **2023 IEEE International Conference on Bioinformatics and Biomedicine (BIBM)**. [S.l.], 2023. p. 4081–4085.

LI, Y.; DAHO, M. E. H.; CONZE, P.-H.; HAJJ, H. A.; BONNIN, S.; REN, H.; MANIVANNAN, N.; MAGAZZENI, S.; TADAYONI, R.; COCHENER, B. et al. Multimodal information fusion for glaucoma and diabetic retinopathy classification. In: SPRINGER. **International Workshop on Ophthalmic Medical Image Analysis**. [S.l.], 2022. p. 53–62.

LIU, Z.; HU, Y.; QIU, Z.; NIU, Y.; ZHOU, D.; LI, X.; SHEN, J.; JIANG, H.; LI, H.; LIU, J. Cross-modal attention network for retinal disease classification based on multi-modal images. **Biomedical Optics Express**, Optica Publishing Group, v. 15, n. 6, p. 3699–3714, 2024.

LUCENTE, A.; TALONI, A.; SCORCIA, V.; GIANNACCARE, G. Widefield and ultra-widefield retinal imaging: A geometrical analysis. **Life**, MDPI, v. 13, n. 1, p. 202, 2023.

MA, J.; LV, B.; LI, Y.; FAN, P.; ZHAO, X.; YUAN, H.; ZHANG, Y. Multimodal primary open angle glaucoma early diagnosing program based on clinical process. 2021.

MCHUGH, M. Interrater reliability: The kappa statistic. **Biochemia medica : časopis Hrvatskoga društva medicinskih biokemičara / HDMB**, v. 22, p. 276–82, 10 2012.

MEHTA, P.; PETERSEN, C. A.; WEN, J. C.; BANITT, M. R.; CHEN, P. P.; BOJIKIAN, K. D.; EGAN, C.; LEE, S.-I.; BALAZINSKA, M.; LEE, A. Y. et al. Automated detection of glaucoma with interpretable machine learning using clinical data and multimodal retinal images. **American Journal of Ophthalmology**, Elsevier, v. 231, p. 154–169, 2021.

MWANZA, J.-C.; OAKLEY, J. D.; BUDENZ, D. L.; CHANG, R. T.; KNIGHT, O. J.; FEUER, W. J. Macular ganglion cell–inner plexiform layer: Automated detection and thickness reproducibility with spectral domain–optical coherence tomography in glaucoma. **Investigative Ophthalmology and Visual Science**, v. 52, n. 11, p. 8323–8329, 10 2011. ISSN 1552-5783. Disponível em: <https://doi.org/10.1167/iovs.11-7962>.

NGUYEN, L. D.; GAO, R.; LIN, D.; LIN, Z. Biomedical image classification based on a feature concatenation and ensemble of deep cnns. **Journal of Ambient Intelligence and Humanized Computing**, Springer, p. 1–13, 2019.

PAWŁOWSKI, M.; WRÓBLEWSKA, A.; SYSKO-ROMAŃCZUK, S. Effective techniques for multimodal data fusion: A comparative analysis. **Sensors**, MDPI, v. 23, n. 5, p. 2381, 2023.

RABIOLO, A.; PARRAVANO, M.; QUERQUES, L.; CICINELLI, M. V.; CARNEVALI, A.; SACCONI, R.; CENTODUCATI, T.; VUJOSEVIC, S.; BANDELLO, F.; QUERQUES, G. Ultra-wide-field fluorescein angiography in diabetic retinopathy: a narrative review. **Clinical Ophthalmology**, Taylor & Francis, p. 803–807, 2017.

RONNEBERGER, O.; FISCHER, P.; BROX, T. U-net: Convolutional networks for biomedical image segmentation. In: SPRINGER. **International Conference on Medical image computing and computer-assisted intervention**. [S.l.], 2015. p. 234–241.

RUIA, S.; TRIPATHY, K. Optical coherence tomography in diabetic retinopathy. In: **Diabetic Eye Disease-From Therapeutic Pipeline to the Real World**. [S.l.]: IntechOpen, 2021.

RUSSAKOVSKY, O.; DENG, J.; SU, H.; KRAUSE, J.; SATHEESH, S.; MA, S.; HUANG, Z.; KARPATHY, A.; KHOSLA, A.; BERNSTEIN, M. et al. Imagenet large scale visual recognition challenge. **International journal of computer vision**, Springer, v. 115, n. 3, p. 211–252, 2015.

SAGI, O.; ROKACH, L. Ensemble learning: A survey. **Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery**, Wiley Online Library, v. 8, n. 4, p. e1249, 2018.

SARHAN, A.; ROKNE, J.; ALHAJJ, R. Glaucoma detection using image processing techniques: A literature review. **Computerized Medical Imaging and Graphics**, Elsevier, p. 101657, 2019.

SBG. 3º consenso de glaucoma primário de Ângulo aberto. Sociedade Brasileira de Glaucoma, 2009. <https://www.sbglaucoma.org.br/wp-content/uploads/2020/06/consenso03-v2.pdf>. Acessado em: 12-Janeiro-2022.

SBG. 2º consenso de glaucoma primário de Ângulo fechado. Sociedade Brasileira de Glaucoma, 2012. [Acessado em 12-Janeiro-2021]. Disponível em: <https://www.sbglaucoma.org.br/wp-content/uploads/2020/06/consenso04-v2.pdf>.

SIMONYAN, K.; ZISSERMAN, A. Very deep convolutional networks for large-scale image recognition. **arXiv preprint arXiv:1409.1556**, 2014.

SOH, Z.; YU, M.; BETZLER, B. K.; MAJITHIA, S.; THAKUR, S.; THAM, Y. C.; WONG, T. Y.; AUNG, T.; FRIEDMAN, D. S.; CHENG, C.-Y. The global extent of undetected glaucoma in adults: a systematic review and meta-analysis. **Ophthalmology**, Elsevier, v. 128, n. 10, p. 1393–1404, 2021.

SOLIMAN, A. Z.; SILVA, P. S.; AIELLO, L. P.; SUN, J. K. Ultra-wide field retinal imaging in detection, classification, and management of diabetic retinopathy. In: TAYLOR & FRANCIS. **Seminars in ophthalmology**. [S.l.], 2012. v. 27, n. 5-6, p. 221–227.

SOLOVYEV, R.; KALININ, A. A.; GABRUSEVA, T. 3d convolutional neural networks for stalled brain capillary detection. **Computers in Biology and Medicine**, Elsevier, v. 141, p. 105089, 2022.

SOUZA, T. R. de; PASCHOINI, L. A. K.; SILLOS, I. R.; MACHADO, E.; SILVA, M. P. B. Manifestações clínicas do glaucoma: Uma revisão narrativa de literatura. **Revista Ibero-Americana de Humanidades, Ciências e Educação**, v. 9, n. 9, p. 813–819, 2023.

STUDY, V. L. E. G. of the Global Burden of D. et al. Global estimates on the number of people blind or visually impaired by glaucoma: A meta-analysis from 2000 to 2020. **Eye**, v. 38, n. 11, p. 2036, 2024.

SUCIU, C.-I.; SUCIU, V.-I.; NICOARA, S.-D. et al. Optical coherence tomography (angiography) biomarkers in the assessment and monitoring of diabetic macular edema. **Journal of Diabetes Research**, Hindawi, v. 2020, 2020.

SZEGEDY, C.; LIU, W.; JIA, Y.; SERMANET, P.; REED, S.; ANGUELOV, D.; ERHAN, D.; VANHOUCKE, V.; RABINOVICH, A. Going deeper with convolutions. In: **Proceedings of the IEEE conference on computer vision and pattern recognition**. [S.l.: s.n.], 2015. p. 1–9.

TAN, T.-E.; WONG, T. Y. Diabetic retinopathy: Looking forward to 2030. **Frontiers in Endocrinology**, Frontiers, v. 13, p. 1077669, 2023.

THAM, Y.-C.; LI, X.; WONG, T. Y.; QUIGLEY, H. A.; AUNG, T.; CHENG, C.-Y. Global prevalence of glaucoma and projections of glaucoma burden through 2040: a systematic review and meta-analysis. **Ophthalmology**, Elsevier, v. 121, n. 11, p. 2081–2090, 2014.

TIAN, H.; LU, S.; SUN, Y.; LI, H. Gc-net: Global and class attention blocks for automated glaucoma classification. In: **2022 IEEE 17th Conference on Industrial Electronics and Applications (ICIEA)**. [S.l.: s.n.], 2022. p. 498–503.

WANG, X.; WANG, M.; LIU, H.; MERCIECA, K.; PRINZ, J.; FENG, Y.; PROKOSCH, V. The association between vascular abnormalities and glaucoma—what comes first? **International Journal of Molecular Sciences**, MDPI, v. 24, n. 17, p. 13211, 2023.

WANG, Y.; ZHEN, L.; TAN, T.-E.; FU, H.; FENG, Y.; WANG, Z.; XU, X.; GOH, R. S. M.; NG, Y.; CALHOUN, C. et al. Geometric correspondence-based multimodal learning for ophthalmic image analysis. **IEEE Transactions on Medical Imaging**, IEEE, 2024.

WANG, Z.; WANG, J.; ZHANG, H.; YAN, C.; WANG, X.; WEN, X. Mstnet: method for glaucoma grading based on multimodal feature fusion of spatial relations. **Physics in Medicine & Biology**, IOP Publishing, v. 68, n. 24, p. 245002, 2023.

WHO. World report on vision. World Health Organization, 2019. <https://www.who.int/publications/i/item/9789241516570>.

WOO, S.; PARK, J.; LEE, J.-Y.; KWEON, I. S. Cbam: Convolutional block attention module. In: **Proceedings of the European conference on computer vision (ECCV)**. [S.l.: s.n.], 2018. p. 3–19.

WU, J.; FANG, H.; LI, F.; FU, H.; LIN, F.; LI, J.; HUANG, L.; YU, Q.; SONG, S.; XU, X. et al. Gamma challenge: Glaucoma grading from multi-modality images. **arXiv preprint arXiv:2202.06511**, 2022.

WU, Z.; XUE, B.; ZHANG, M. Multitree gp-based feature learning for multimodal medical image classification. In: IEEE. **2025 IEEE Congress on Evolutionary Computation (CEC)**. [S.l.], 2025. p. 1–8.

XIONG, J.; LI, F.; SONG, D.; TANG, G.; HE, J.; GAO, K.; ZHANG, H.; CHENG, W.; SONG, Y.; LIN, F. et al. Multimodal machine learning using visual fields and peripapillary circular oct scans in detection of glaucomatous optic neuropathy. **Ophthalmology**, Elsevier, 2021.

YU, Y.; ZHU, H.; QIAN, T.; CHEN, N.; HUANG, B. Modality-specificity multi-aware evidence fusion algorithm using cfp and oct for fundus diseases diagnosis. **Pattern Recognition**, Elsevier, p. 111957, 2025.

YU, Y.; ZHU, H.; QIAN, T.; HOU, T.; HUANG, B. Multi-task collaboration for cross-modal generation and multi-modal ophthalmic diseases diagnosis. **IET Image Processing**, Wiley Online Library, v. 19, n. 1, p. e70016, 2025.

ZHANG, X.; FRANCIS, B. A.; DASTIRIDOU, A.; CHOPRA, V.; TAN, O.; VARMA, R.; GREENFIELD, D. S.; SCHUMAN, J. S.; HUANG, D.; GROUP, A. I. for G. S. et al. Longitudinal and cross-sectional analyses of age effects on retinal nerve fiber layer and ganglion cell complex thickness by fourier-domain oct. **Translational vision science & technology**, The Association for Research in Vision and Ophthalmology, v. 5, n. 2, p. 1–1, 2016.

ZHAO, J.; LI, S.; HAO, Y.; ZHANG, C. Bridging the modality gap in multimodal eye disease screening: learning modality shared-specific features via multi-level regularization. **IEEE Signal Processing Letters**, IEEE, 2025.

ZOU, K.; LIN, T.; HAN, Z.; WANG, M.; YUAN, X.; CHEN, H.; ZHANG, C.; SHEN, X.; FU, H. Confidence-aware multi-modality learning for eye disease screening. **Medical Image Analysis**, Elsevier, v. 96, p. 103214, 2024.