UNIVERSIDADE FEDERAL DE OURO PRETO
INSTITUTO DE CIÊNCIAS EXATAS E BIOLÓGICAS
DEPARTAMENTO DE COMPUTAÇÃO

ARTHUR HENRIQUE SANTOS CELESTINO

CLASSIFICAÇÃO DE GLAUCOMA EM IMAGENS DE FUNDO DE OLHO

Ouro Preto, MG
2025

ARTHUR HENRIQUE SANTOS CELESTINO

CLASSIFICAÇÃO DE GLAUCOMA EM IMAGENS DE FUNDO DE OLHO

Monografia apresentada ao Curso de Ciência da Computação da Universidade Federal de Ouro Preto como parte dos requisitos necessários para a obtenção do grau de Bacharel em Ciência da Computação.

Orientador: Pedro Henrique Lopes Silva

Ouro Preto, MG
2025

SISBIN - SISTEMA DE BIBLIOTECAS E INFORMAÇÃO

C392c Celestino, Arthur Henrique Santos.
Classificação de glaucoma em imagens de fundo de olho. [manuscrito] / Arthur Henrique Santos Celestino. - 2025.
61 f.: il.: color., gráf., tab..

Orientador: Prof. Dr. Pedro Henrique Lopes Silva.
Monografia (Bacharelado). Universidade Federal de Ouro Preto. Instituto de Ciências Exatas e Biológicas. Graduação em Ciência da Computação .

1. Inteligência artificial. 2. Glaucoma. 3. Redes neurais convolucionais. I. Silva, Pedro Henrique Lopes. II. Universidade Federal de Ouro Preto. III. Título.

CDU 004.8:617.7

Bibliotecário(a) Responsável: Sione Galvão Rodrigues - CRB6 / 2526

MINISTÉRIO DA EDUCAÇÃO
UNIVERSIDADE FEDERAL DE OURO PRETO
REITORIA
INSTITUTO DE CIENCIAS EXATAS E BIOLOGICAS
DEPARTAMENTO DE COMPUTACAO

FOLHA DE APROVAÇÃO

Arthur Henrique Santos Celestino

Classificação de Glaucoma em Imagens de fundo de olho

Monografia apresentada ao Curso de Ciência da Computação da Universidade Federal de Ouro Preto como requisito parcial para obtenção do título de Bacharel em Ciência da Computação

Aprovada em 27 de Agosto de 2025.

Membros da banca

Pedro Henrique Lopes Silva (Orientador) - Doutor - Universidade Federal de Ouro Preto
Guilherme Augusto Anício Drummond do Nascimento (Examinador) - Bacharel - Universidade Federal de Ouro Preto
Luan Patrik Silva Pinto (Examinador) - Bacharel - Programa de Pós-Graduação em Ciência da Computação - UFOP

Pedro Henrique Lopes Silva, Orientador do trabalho, aprovou a versão final e autorizou seu depósito na Biblioteca Digital de Trabalhos de Conclusão de Curso da UFOP em 27/08/2025.

Documento assinado eletronicamente por **Pedro Henrique Lopes Silva**, **PROFESSOR DE MAGISTERIO SUPERIOR**, em 27/08/2025, às 11:05, conforme horário oficial de Brasília, com fundamento no art. 6º, § 1º, do Decreto nº 8.539, de 8 de outubro de 2015.

A autenticidade deste documento pode ser conferida no site http://sei.ufop.br/sei/controlador_externo.php?acao=documento_conferir&id_orgao_acesso_externo=0 , informando o código verificador **0965546** e o código CRC **857FAA17**.

R. Diogo de Vasconcelos, 122, - Bairro Pilar Ouro Preto/MG, CEP 35402-163
Telefone: 3135591692 - www.ufop.br

Dedico esse trabalho à todos meus amigos e familiares que foram pilares em minha jornada, em especial aos meus pais Josie Celestino e Edivania dos Santos Reis Celestino, além da minha avó Edir Adriana.

Agradecimentos

Gostaria, primeiramente, de agradecer aos meus familiares, em especial aos meus pais, Josie e Edivania, por todos os ensinamentos e por todo sacrifício que me trouxe até onde estou. Passei toda a minha vida vendo toda a luta deles para que eu pudesse, enfim, chegar a esse momento; por isso, o meu mais sincero agradecimento. Também gostaria de agradecer, em especial, a minha avó, dona Edir, por sempre ser alguém que me ofereceu tanto carinho, mesmo nos momentos em que estive mais distante. Amo muito todos vocês.

Após a família de sangue, gostaria de agradecer a minha segunda família em Ouro Preto: a República Badalação. Mesmo após divergências de caminhos, todos sempre foram um apoio muito forte nos momentos difíceis, mesmo sem conhecimento. Por vezes, em meio ao caos e a tantas incertezas, era um dos poucos lugares que eu conseguia aliviar a tensão, e simplesmente conversar e rir de qualquer coisa. Realmente foi a minha segunda casa, por isso, agradeço muito. Ainda no ambiente republicano, também deixo um agradecimento às meninas da República Água na Boca, por sempre terem sido tão atenciosas comigo, sempre se mantendo como verdadeiras amigas.

Ao meu grupo de amigos, grupo dos canalhas, composto por mim, Mauro, Marcella e Silas. Quem diria que de um grupo que entrava no *discord* para se ajudar na faculdade no primeiro período, terminaríamos como esse grupo que sempre esteve junto, moramos juntos, e dividimos partes de nossas vidas. Levo nosso grupo em um lugar especial no meu coração.

Por fim, agradeço aos meus professores, em especial ao meu orientador, Pedro Henrique Lopes Silva, por todo o suporte nessa trajetória, por aguentar meu grande otimismo e ânimo, e por sempre me puxar para frente, além de, obviamente, por todo apoio na monografia. Esse trabalho só está sendo possível por você.

“Os gênios vivem apenas uma história de loucura.”

(SCHOPENHAUER, 2021)

Resumo

O glaucoma é uma das principais causas de cegueira irreversível no mundo, caracterizado pelo dano progressivo ao nervo óptico. O diagnóstico precoce é essencial para minimizar os impactos da doença, e a Inteligência Artificial tem se mostrado uma ferramenta importante nesse contexto. Este trabalho apresenta uma análise exploratória de diferentes arquiteturas de redes neurais, incluindo redes convolucionais e modelos baseados em *transformers*, para a classificação de imagens de fundo de olho com glaucoma, utilizando a base de dados *Standardized Multi-Channel Dataset for Glaucoma*. Um dos focos principais é avaliar o impacto da combinação das previsões através de esquemas de votação para mitigar erros de modelos individuais e aprimorar o desempenho geral. Inicialmente, foram avaliados os modelos pré-treinados ResNet-50, VGG-19 e EfficientNet-B0, além do *Vision Transformer* (ViT) e MaxViT. Os experimentos demonstraram que a ResNet-50 (acurácia de 87,23%) e a VGG-19 (acurácia de 86,04%) obtiveram resultados consistentes, mas com *recall* inferior. A EfficientNet-B0 apresentou desempenho significativamente menor (acurácia de 69,96%), sendo considerada demasiadamente simples para a tarefa. Entre os *transformers*, o MaxViT destacou-se com o maior *recall* (93,20%), característica crucial para a detecção de casos positivos. O ViT obteve um desempenho mais equilibrado (acurácia de 87,99%, recall de 82,44%).Os esquemas de votação, especialmente a votação por média, demonstraram ser uma abordagem com um desempenho geral superior e mais equilibrado, com acurácia de 89,45%.

Palavras-chave: Glaucoma. Redes Neurais Convolucionais. *Transformers*. Votação.

Abstract

Glaucoma is one of the leading causes of irreversible blindness worldwide, characterized by the progressive damage to the optic nerve. Early diagnosis is essential to minimize the disease's impact, and Artificial Intelligence has proven to be an important tool in this context. This work presents an exploratory analysis of different neural network architectures, including convolutional networks and transformer-based models, for the classification of fundus images with glaucoma using the *Standardized Multi-Channel Dataset for Glaucoma* dataset. One of the main focuses is to evaluate the impact of combining predictions through voting schemes in order to mitigate individual model errors and improve overall performance. Initially, pre-trained models such as ResNet-50, VGG-19, and EfficientNet-B0 were assessed, along with *Vision Transformer* (ViT) and MaxViT. The experiments showed that ResNet-50 (accuracy of 87.23%) and VGG-19 (accuracy of 86.04%) achieved consistent results, though with lower recall. EfficientNet-B0 presented significantly lower performance (accuracy of 69.96%), being considered too simple for the task. Among the transformers, MaxViT stood out with the highest recall (93.20%), a crucial characteristic for the detection of positive cases. The ViT achieved a more balanced performance (accuracy of 87.99%, recall of 82.44%). The voting schemes, especially soft voting, demonstrated superior and more balanced overall performance, with an accuracy of 89.45%.

Keywords: Glaucoma. Convolutional Neural Networks. Transformers. Voting.

Lista de Ilustrações

Lista de Tabelas

Lista de Abreviaturas e Siglas

AUC *Area Under the Curve*. 2, 25–27, 41, 44–46, 51

CA-ViT *Contour-Guided and Augmented Vision Transformer*. 28

CLS *Classification Token*. 17

CNN *Convolutional Neural Network*. vii, xiii, xiv, 2, 3, 5, 12, 13, 15, 16, 18, 20, 22, 23, 26–28, 32–34, 42, 44, 47, 52

CVGAN *Conditional Variational Generative Adversarial Network*. 29

DETR *Detection Transformer*. 29

FN *False Negative*. vii, 23, 24, 44, 48, 49

FP *False Positive*. 23, 24, 48, 49

GAN *Generative Adversarial Networks*. 22

HRF *High-Resolution Fundus*. 26

IA Inteligência Artificial. xiii, 2, 6, 7

KNN *K-Nearest Neighbors*. 28

MLP *Multi-Layer Perceptron*. 17

MSE *Mean Squared Error*. 9

OC *Optic Cup*. 29, 32

OCT *Optical Coherence Tomography*. 2

OD *Optic Disc*. 29, 31, 32, 48, 49, 53

ORIGA *Online Retinal Fundus Image Dataset for Glaucoma Analysis and Research*. 27

ReLU *Rectified Linear Unit*. 8, 13, 14, 37

ROC *Receiver Operating Characteristic*. vii, 25

ROI *Region of Interest*. 26

SMDG-19 *Standardized Multi-Channel Dataset for Glaucoma*. v, vi, 2, 29–31, 50, 52

SVM *Support Vector Machine*. 28

TN *True Negative*. 24

TP *True Positive*. 24

ViT *Vision Transformer*. v–ix, xiii, xiv, 2, 5, 16–18, 20, 28, 29, 32, 33, 38, 39, 43–52

Lista de Símbolos

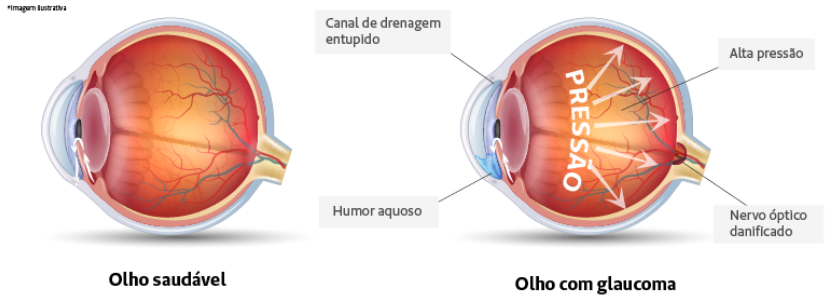
σ	Letra grega minúscula sigma, representando a função sigmoide
α	Letra grega minúscula alpha, representando uma constante positiva
$\sum$	Somatório
η	Letra grega minúscula eta, representando a taxa de aprendizado
∂	Derivada parcial
θ	Letra grega minúscula theta, representando parâmetros do modelo.
∇	Operador nabla, representando gradiente.
γ	Letra grega minúscula gama, representando fator de decaimento exponencial.
ϵ	Letra grega minúscula épsilon, sendo um termo de estabilidade numérica.
β	Letra grega minúscula beta para fatores de decaimento.
λ	Letra grega minúscula lambda, representando coeficiente de *weight decay*.
$\in$	Símbolo de pertinência a um conjunto.
$\hat{\cdot}$	Acento circunflexo “hat” para estimativas corrigidas de viés.
$\bar{\cdot}$	Barra superior indicando média.

Sumário

1 Introdução

O Glaucoma é uma condição que afeta o nervo óptico devido a uma elevação da pressão ocular, como visto na Figura 1.1, a qual pode resultar em danos permanentes e, eventualmente, levar à cegueira (RICARDO, 2024). É uma doença ocular com impacto significativo na saúde pública global, especialmente entre idosos (THAM et al., 2014). Em 2020, estimava-se que 76 milhões de pessoas entre 40 e 80 anos fossem afetadas pela doença, com projeções indicando um aumento para 111,8 milhões até 2040. Dentre esses casos, aproximadamente 71,3% devem ser classificados como glaucoma primário de ângulo aberto, sendo o maior crescimento esperado entre populações da Ásia e da África (ALLISON et al., 2024).

Figura 1.1 – Aumento da pressão ocular que leva ao glaucoma.



Fonte: (Unimed, 2025).

A enfermidade acontece quando o líquido responsável pela nutrição e proteção ocular não consegue ser drenado adequadamente, levando ao acúmulo e ao aumento da pressão intraocular. Esse aumento pressiona e danifica gradualmente o nervo óptico, estrutura essencial para a comunicação entre os olhos e o cérebro, resultando em prejuízos à visão. Como a degeneração do nervo acontece de forma lenta e silenciosa, os sintomas costumam surgir apenas em estágios mais avançados da doença, geralmente após os 40 anos. Um dos primeiros sinais é a perda da visão periférica, fazendo com que o paciente enxergue apenas o centro do campo visual. Porém, caso não seja tratado rapidamente, o glaucoma pode evoluir para um comprometimento total da visão, culminando na cegueira (SENIOR, 2023).

A detecção do glaucoma é realizada por profissionais da saúde ocular, como oftalmologistas e optometristas, por meio de uma série de exames clínicos e diagnósticos, seja por imagens ou não (ALLISON et al., 2024). No ramo das técnicas mais tradicionais, a medição da pressão intraocular é um dos principais indicadores (uma vez que ela está ligada ao desenvolvimento da doença em si), podendo ser medida por métodos como tonometria. O exame do nervo óptico é essencial para detectar danos estruturais, geralmente avaliado através da relação escavação/disco óptico. A gonioscopia também é realizada para analisar o ângulo iridocorneano e classificar o tipo

de glaucoma (WEINREB; AUNG; MEDEIROS, 2014). Esses exames, combinados, permitem um diagnóstico mais preciso e auxiliam na definição do tratamento adequado para prevenir a progressão da doença.

Porém, esses exames acabam se mostrando uma alternativa custosa tanto financeiramente quanto em relação ao tempo para obtenção do resultado (KRIžAJ, 2019). Outras alternativas são exames de imagem, que incluem fotografia do disco óptico e *Optical Coherence Tomography* (OCT), empregando interferometria para obter imagens detalhadas da retina e analisar deformações estruturais causadas pelo glaucoma (UR;, 2019). Apesar de eficazes, essas técnicas (tanto de imagem quanto as baseadas em medições) apresentam restrições em sensibilidade e precisão. A OCT, por exemplo, é eficaz na identificação de mudanças estruturais nas camadas da retina, porém não consegue detectar precocemente a perda funcional. Já o teste de campo visual mede o impacto funcional da doença, mas tem baixa precisão na identificação de alterações estruturais. Além disso, a grande variação na manifestação do glaucoma, que pode apresentar sinais discretos no início e danos severos em estágios avançados, torna o diagnóstico mais complexo (ASHTARI-MAJLAN; DEHSHIBI; MASIP, 2023).

Uma resposta a esses desafios é o uso de Inteligência Artificial (IA), potencializada pelo crescimento das redes de aprendizado profundo, também conhecidas como *Deep Learning* e principalmente das *Convolutional Neural Network* (CNN), utilizadas para predição e diagnóstico de casos de glaucoma (OH et al., 2021). Nesse contexto, além de abordagens mais tradicionais envolvendo redes convolucionais (SAXENA et al., 2020), alguns trabalhos (CHEN et al., 2015; VELPULA et al., 2024) apresentaram resultados utilizando estratégias com CNNs, enquanto outros se basearam na utilização de um sistema de votação majoritária entre redes distintas (VELPULA; SHARMA, 2023), e até mesmo a utilização de *Vision Transformer* (ViT), modelo este bem mais complexo e, consequentemente, custoso, para a classificação (DOSOVITSKIY et al., 2021; YURDAKUL; UYAR; TASDEMIR, 2025).

A pergunta de pesquisa deste trabalho é: *Utilizar a votação entre modelos diversos pode impactar positivamente o desempenho geral?*. Dessa forma, é possível identificar um modelo que ofereça uma solução robusta em uma base de dados mais complexa, como a *Standardized Multi-Channel Dataset for Glaucoma* (SMDG-19), que ainda carece de uma base maior de trabalhos.

Os resultados obtidos individualmente (por exemplo, ResNet-50 com aproximadamente 87% de acurácia e MaxViT com 93% de *recall*), e, principalmente, através dos esquemas de votação, demonstraram que a combinação das previsões pode sim levar a um desempenho geral superior e mais equilibrado, como observado com a votação que se utiliza das médias das probabilidades, alcançando 89,45% de acurácia e 0,9567 de *Area Under the Curve* (AUC). Esses achados solidificam a base para futuras otimizações, especialmente no que tange à integração com técnicas avançadas de pré-processamento, atenção e generalização.

1.1 Justificativa

O Glaucoma é uma doença extremamente preocupante devido à sua dificuldade de detecção em fases iniciais. A sua detecção precoce é essencial para evitar danos permanentes na visão, como a cegueira (ASHTARI-MAJLAN; DEHSHIBI; MASIP, 2023).

A escolha do tema também se dá pela busca de um modelo capaz de classificar imagens de fundo de olho com alta precisão e confiabilidade, ao mesmo tempo que não necessita de pré-processamentos complexos. Estudos anteriores (ARAúJO et al., 2017) indicam que modelos com acurácia superior a 88% e *recall* acima de 84% são considerados eficazes para a detecção de glaucoma. Dessa forma, busca-se desenvolver uma abordagem que atinja ou supere esses níveis de desempenho descritos, tornando-a mais prática e de mais fácil integração em sistemas de diagnóstico, evitando etapas intermediárias complexas (como segmentação para ser usada por outra rede). Além disso, um dos principais pontos motivadores é a utilização de uma base maior de dados em estratégias já conhecidas, procurando averiguar o quanto o estado atual dos dados na área limita os avanços.

1.2 Objetivos

O objetivo principal do trabalho é construir um conjunto de modelos baseados em redes neurais convolucionais (CNNs) e *transformers* para classificação de imagens de fundo de olho em dois grupos principais: as que indicam glaucoma e as que não possuem a doença. Porém, o principal foco se faz presente na junção destes modelos no momento da classificação, visando assim, com a combinação das previsões, mitigar erros que modelos individuais poderiam cometer, aproveitando as diferentes capacidades de cada arquitetura em capturar padrões nas imagens de fundo de olho. Dessa forma, busca-se alcançar um resultado que possa ser útil no diagnóstico prático da doença.

Tendo em vista o objetivo principal, pode-se traçar alguns intermediários, como:

- **Avaliação de modelos distintos de classificação**: Nesta etapa, tem-se como objetivo selecionar alguns modelos para o processo de classificação das imagens, sejam eles modelos fundadores para as CNNs ou modelos pré-treinados de *transformers*.
- **Avaliação dos diferentes modelos individualmente**: Nesta etapa, serão avaliados diferentes variações nas redes CNNs utilizadas, além de também se olhar para o desempenho dos modelos *transformers*, cada um individualmente.
- **Desenvolvimento da predição combinando os modelos**: A etapa que envolve a criação do algoritmo que irá realizar a classificação em si dos dados, após já treinados os modelos prévios, dando o diagnóstico do glaucoma. Espera-se que seja um resultado superior ao individual de cada um.

- **Avaliação do desempenho**: Na fase de analisar o desempenho, serão calculadas métricas comuns como acurácia, revocação (também conhecido como *recall*), precisão, F1-Score e AUC quando possível, permitindo comparação fácil com a literatura.
- **Comparação do resultado**: Após avaliar individualmente os resultados obtidos, é importante analisar-se a relevância do trabalho perante ao que já existe na literatura, procurando analisar o balanço entre desempenho obtido e complexidade da solução.

1.3 Organização do Trabalho

O Capítulo 2 oferece uma revisão da literatura relacionada ao tema, abordando tanto os estudos que aplicam diferentes técnicas de aprendizado para classificação de glaucoma quanto a base teórica necessária para compreender o estudo desenvolvido. O Capítulo 3 detalha a metodologia adotada, descrevendo as etapas do desenvolvimento e a implementação dos modelos propostos. No Capítulo 4, são apresentados e discutidos os resultados dos experimentos conduzidos. Por fim, no Capítulo 5 estão as conclusões finais e possíveis ações futuras.

2 Revisão Bibliográfica

Este capítulo contextualiza os principais tópicos necessários para o devido entendimento da pesquisa realizada. Na Seção 2.1 é fornecida uma base teórica utilizada para a compreensão dos conceitos e metodologias empregados no presente estudo. Já na Seção 2.2 são apresentados os trabalhos relacionados, com o objetivo de apresentar o que já foi proposto e testado em relação à classificação de glaucoma em imagens de fundo de olho.

2.1 Fundamentação Teórica

Neste capítulo, serão apresentados os conceitos teóricos necessários para o entendimento do estudo. Inicialmente, será fornecida uma visão geral sobre o campo de *Machine Learning*, que é bastante amplo. Em seguida, o foco será direcionado às redes neurais artificiais, com ênfase nas CNNs e ViTs, abordando os principais conceitos e camadas utilizadas, além de uma visão geral do processo de aprendizado. Por fim, será discutido o conceito de *transfer learning*, técnica que será empregada no desenvolvimento deste trabalho, visando maior eficiência.

2.1.1 Glaucoma

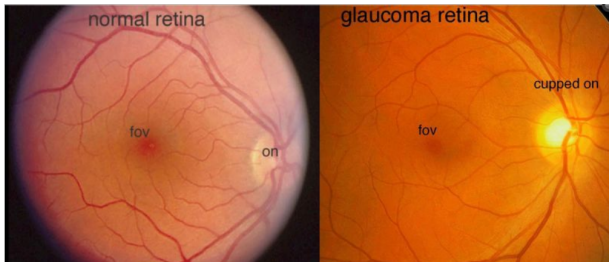
O Glaucoma é uma das principais causas de cegueira no mundo devido à perda irreversível dos neurônios da retina, com grandes dificuldades para detecção em estágios iniciais. Embora seja comumente referenciado como uma única doença, o termo, na verdade, se refere a um grupo heterogêneo de doenças caracterizadas por alterações biomecânicas no olho. Atingindo milhões de pessoas em todo o mundo, a doença provoca um impacto econômico substancial, tanto pelos custos diretos relacionados aos tratamentos médicos quanto pelos efeitos indiretos, como a diminuição da produtividade e a piora na qualidade de vida (KRIžAJ, 2019).

Vale ressaltar que os critérios atuais para o diagnóstico do glaucoma ainda não incluem a identificação de marcadores moleculares e fisiológicos da doença. Em outras palavras, ainda não há uma estrutura que relacione os diversos fatores de risco e fenótipos do glaucoma, o que constitui uma limitação significativa no diagnóstico e no tratamento. A definição tradicional do glaucoma como “uma neuropatia óptica caracterizada pela ‘escavação’ da cabeça do nervo óptico” é considerada ultrapassada, pois descreve um estágio avançado da condição da doença (KRIžAJ, 2019). A atual definição discute o fato de que o glaucoma pode estar presente antes mesmo de alterações significativas na cabeça do nervo óptico serem detectadas, focando na necessidade de identificar sinais precoces e alterações sutis para um diagnóstico mais proativo e preciso.

O diagnóstico do glaucoma, no contexto clínico, é realizado por meio da avaliação da pressão intraocular (que pode não ser muito conclusiva devido a variações naturais (QUIGLEY,

1993)) e da análise visual de exames diversos voltados para a análise do nervo óptico. Na Figura 2.1 é possível ver um exemplo de comparação realizado nesse aspecto. Esse tipo de análise acaba levando a altos custos e, principalmente, a um grande consumo de tempo, devido à minúcia necessária, levando a uma busca por alternativas como o aprendizado de máquina.

Figura 2.1 – Comparação de uma imagem de fundo de olho com glaucoma (a direita), contra uma normal (esquerda).



Fonte: (KRIžAJ, 2019).

2.1.2 Inteligência Artificial e *Machine Learning*

A IA é um campo de estudo que busca criar maneiras de resolver problemas por meio de sistemas que simulem o conhecimento humano (MITCHELL, 1997). Este campo abrange uma variedade de subdisciplinas, com o objetivo de criar máquinas que possam aprender, raciocinar e interagir de maneira autônoma com o ambiente. A evolução da IA tem sido marcada por avanços significativos, desde os primeiros programas de resolução de problemas até as complexas arquiteturas de inteligência incorporada que buscam replicar aspectos da cognição humana (BRUNETTE; FLEMMER; FLEMMER, 2009). Ela se encontra cada vez mais presente em diversos contextos da sociedade, seja para comodidade, lazer e até mesmo em diagnósticos na área médica como em (TAJ et al., 2021), (BRAGANÇA; TORRES; SOARES, 2023) e (VELPULA; SHARMA, 2023).

Uma das principais subáreas da IA é *Machine Learning*. Nela são desenvolvidos algoritmos e modelos estatísticos que permitem aos computadores realizar tarefas de forma autônoma, identificando padrões em grandes volumes de dados históricos. Esses algoritmos ajudam a fazer previsões precisas com base em dados de entrada. Por exemplo, é possível treinar um sistema médico para diagnosticar doenças a partir de exames, usando imagens digitalizadas e seus diagnósticos associados (MITCHELL, 1997), como é o caso do glaucoma. De maneira básica, os algoritmos funcionam seguindo um fluxo de treinamento utilizando dados já conhecidos e assumidamente verdadeiros (chamados de *ground-truth*), formando assim um modelo para tentar aproximar a resposta a novos dados que forem apresentados ao mesmo (ZHOU, 2021). Porém, esta é apenas uma visão geral do processo, já que ele pode se apresentar de diferentes formas a variar da disponibilidade de dados e do objetivo requerido.

As diferentes técnicas de *Machine Learning* podem ser categorizadas em quatro categorias principais: (i) o aprendizado supervisionado, (ii) semi-supervisionado, (iii) não supervisionado, e

(iv) por reforço. No aprendizado supervisionado, um dos mais comuns, são utilizados no processo de treinamento rótulos confiáveis junto aos dados, dessa forma avaliando as correlações para futuramente dar previsões semelhantes a novos dados (JANIESCH; ZSCHECH; HEINRICH, 2021), podendo ser discretas (como modelos de decisão se um e-mail é spam ou não) ou contínuas (como em predições financeiras).

Já no não supervisionado, existem situações onde o sistema precisa identificar padrões nos dados sem contar com rótulos ou especificações previamente definidos. Nesse caso, o conjunto de dados de treinamento contém apenas dados brutos, e o objetivo é descobrir informações estruturais relevantes, como identificar grupos de elementos com características semelhantes (conhecido como *clustering*). Um exemplo comum desse tipo de aprendizado em mercados eletrônicos é o uso de técnicas de agrupamento para segmentar clientes ou mercados, possibilitando uma comunicação mais personalizada e focada em grupos específicos (JANIESCH; ZSCHECH; HEINRICH, 2021).

O aprendizado semi-supervisionado é uma combinação dos dois anteriores, se mostrando uma alternativa em cenários onde a obtenção de dados e variáveis de interesse representa um processo custoso e demorado, como imagens de certos exames na área da saúde. Nela, têm-se apenas algumas instâncias rotuladas, e apoiando-se nelas, o algoritmo tenta agrupar as restantes (ALMEIDA, 2023).

Por fim, no aprendizado por reforço, em vez de apresentar pares de entrada e saída, o sistema é configurado com uma descrição do estado atual, um objetivo a ser alcançado, uma lista de ações permitidas e as restrições associadas aos resultados dessas ações como recompensas e penalidades. O modelo, então, aprende a atingir o objetivo por meio de tentativa e erro, buscando maximizar uma recompensa. Esse tipo de aprendizado tem sido altamente eficaz em ambientes como jogos e no mercado financeiro (JANIESCH; ZSCHECH; HEINRICH, 2021).

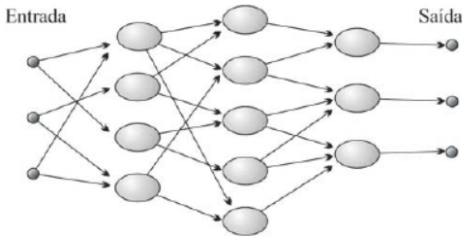
Dentre as diversas abordagens de aprendizado de máquina, algumas técnicas têm se destacado por sua capacidade de modelar representações complexas dos dados de forma hierárquica e automática. Essa evolução levou ao desenvolvimento de arquiteturas mais profundas e especializadas, capazes de extrair padrões sutis e aprimorar o desempenho em diversas tarefas.

2.1.3 *Deep Learning*

Uma subárea ainda mais específica de IA e *Machine Learning* é *Deep Learning*, composta principalmente pelas chamadas redes neurais profundas. Primeiramente, é importante salientar que as redes neurais, no geral, são estruturas que buscam simular os neurônios cerebrais humanos, através de vários nós organizados em camadas distintas, totalmente interligadas entre si, como demonstrado na Figura 2.2. A cada ligação é atribuído um peso que, ao ser aplicado aos dados de entrada e propagado para as camadas subsequentes, busca estabelecer relações numéricas para reconhecer padrões.

Porém, o que diferencia os modelos de *Deep Learning* das redes neurais tradicionais é o uso de três ou mais camadas computacionais, contrastando com os modelos clássicos que utilizam apenas uma ou duas. Essa profundidade adicional possibilita que a rede extraia automaticamente características mais abstratas dos dados, tornando-se uma ferramenta essencial em cenários onde os padrões são sutis ou desconhecidos. No entanto, essa complexidade também traz desafios, como a interpretabilidade dos modelos e a necessidade de maior poder computacional (LECUN; BENGIO; HINTON, 2015).

Figura 2.2 – Arquitetura de uma rede neural básica



Fonte: (FERNEDA, 2006).

O funcionamento de uma rede neural profunda ocorre em duas principais fases: **Feed-forward** e **Backpropagation**.

2.1.3.1 Feedforward

No processo de *Feedforward*, os dados de entrada percorrem a rede, camada por camada, até alcançar a camada de saída. Cada neurônio recebe um conjunto de entradas ponderadas, realiza uma soma e aplica uma função de ativação para introduzir não-linearidade ao modelo. Matematicamente, essa operação pode ser descrita como:

$$z^{(l)} = W^{(l)} a^{(l-1)} + b^{(l)} \tag{2.1}$$

$$a^{(l)} = f(z^{(l)}) \tag{2.2}$$

onde:

- $z^{(l)}$ representa o valor linear combinado antes da ativação na camada l;
- $W^{(l)}$ é a matriz de pesos conectando a camada $l-1$ à camada l;
- $a^{(l-1)}$ são as ativações da camada anterior;
- $b^{(l)}$ é o vetor de vieses associado à camada l;
- $f(z)$ é a função de ativação, como ReLU ou Sigmoide.

Esse processo continua até a última camada, gerando uma saída final (LECUN; BENGIO; HINTON, 2015).

2.1.3.2 Backpropagation

Após calcular a saída da rede, o processo de *Backpropagation* ajusta os pesos para minimizar o erro da predição (ROJAS; ROJAS, 1996). O erro é avaliado por meio de uma função de custo, como o erro quadrático médio (em inglês *Mean Squared Error* - MSE), ou a entropia cruzada, dependendo da tarefa. A atualização dos pesos é feita utilizando o método do gradiente descendente, seguindo a regra:

$$W^{(l)} \leftarrow W^{(l)} - \eta \frac{\partial J}{\partial W^{(l)}}, \tag{2.3}$$

onde:

- J é a função de custo;
- η é a taxa de aprendizado;
- $\frac{\partial J}{\partial W^{(l)}}$ representa o gradiente da função de custo em relação aos pesos.

O *Backpropagation* permite que a rede ajuste seus parâmetros iterativamente para melhorar seu desempenho, tornando-se um dos principais pilares do treinamento de redes neurais profundas (ROJAS; ROJAS, 1996).

2.1.3.3 Funções de custo

Como visto, no *Backpropagation*, é necessário uma função de custo para que o ajuste da rede seja possível. Essas funções são utilizadas para medir o erro entre as previsões do modelo e os valores reais, guiando assim o processo de aprendizado ao fornecer um critério para a otimização dos pesos. A escolha da função de custo depende do tipo de problema, como regressão ou classificação, e pode influenciar diretamente a performance do modelo (AGGARWAL et al., 2018).

As duas mais populares atualmente são as funções *Binary Crossentropy* (apresentada na Equação (2.4) e *Categorical Crossentropy* (apresentada na Equação (2.5)). A primeira é aplicada em problemas de classificação binária, avaliando a distância entre as probabilidades previstas e os rótulos verdadeiros, penalizando previsões erradas (RUBY; YENDAPALLI et al., 2020). Já a segunda é utilizada para classificação multi-classe, utilizadas nos modelos *transformers* deste trabalho, medindo a discrepância entre a distribuição de probabilidade prevista e a verdadeira, incentivando a atribuição de altas probabilidades às classes corretas (ZHANG; SABUNCU, 2018). A função de perda *Binary Crossentropy* é definida por:

$$L = -\frac{1}{N} \sum_{i=1}^{N} \left[y_i \log(\hat{y}_i) + (1 - y_i) \log(1 - \hat{y}_i) \right], \tag{2.4}$$

onde:

- N é o número total de amostras;
- y_i representa o rótulo verdadeiro da i-ésima amostra, assumindo valores 0 ou 1;
- $\hat{y}_i$ é a probabilidade prevista pelo modelo para a i-ésima amostra;
- O logaritmo natural ($\log$) é aplicado para penalizar previsões erradas de forma exponencial.

Já a *Categorical Crossentropy* é definida por:

$$L = -\frac{1}{N}\sum_{i=1}^{N}\sum_{j=1}^{C} y_{ij}\log(\hat{y}_{ij}), \tag{2.5}$$

onde:

- N é o número total de amostras;
- C é o número total de classes;
- y_{ij} representa o rótulo verdadeiro da i-ésima amostra para a j-ésima classe (assume valor 1 para a classe correta e 0 para as demais);
- $\hat{y}_{ij}$ é a probabilidade prevista pelo modelo para a i-ésima amostra pertencer à j-ésima classe;
- O logaritmo natural ($\log$) é aplicado para penalizar previsões erradas de forma exponencial, incentivando o modelo a atribuir maior probabilidade à classe correta.

2.1.3.4 Otimizadores

Otimizadores desempenham um papel crucial no processo de treinamento de redes neurais profundas, agindo durante a fase de *Backpropagation* para ajustar iterativamente os pesos da rede e minimizar o erro de predição. Eles utilizam as informações do gradiente da função de custo para guiar a direção e a magnitude das atualizações dos parâmetros da rede, buscando convergir para um conjunto de pesos que minimize o erro (CHOI et al., 2019).

Existem diversos algoritmos de otimização, e a escolha do otimizador pode impactar significativamente a velocidade de convergência e a performance final do modelo. Uma categoria importante são os otimizadores adaptativos, que ajustam a taxa de aprendizado para cada parâmetro individualmente, com base em estatísticas históricas dos gradientes (ZOU et al., 2019). O RMSProp, cuja formulação pode ser vista na Equação (2.6), é um exemplo de otimizador adaptativo que calcula uma média móvel dos quadrados dos gradientes para normalizar as taxas de aprendizado, tornando o treinamento mais estável e eficiente, especialmente em redes profundas e com gradientes ruidosos (ELSHAMY et al., 2023).

Seja $g_t = \nabla_\theta \mathcal{L}(\theta_t)$ o gradiente da função de perda $\mathcal{L}$ em relação aos parâmetros θ no instante t. O RMSProp mantém uma média móvel exponencial dos quadrados dos gradientes:

$$E[g^2]_t = \gamma E[g^2]_{t-1} + (1-\gamma) g_t^2, \tag{2.6}$$

em que $\gamma \in [0, 1)$ é o fator de decaimento exponencial, que controla a importância relativa dos gradientes passados.

A atualização dos parâmetros é então realizada da seguinte forma:

$$\theta_{t+1} = \theta_t - \frac{\eta}{\sqrt{E[g^2]_t + \epsilon}}\, g_t, \tag{2.7}$$

onde:

- η é a taxa de aprendizado,
- $E[g^2]_t$ é a estimativa da média móvel dos quadrados dos gradientes,
- ϵ é um termo pequeno para garantir estabilidade numérica.

Outro otimizador adaptativo amplamente utilizado, e particularmente eficaz em arquiteturas complexas como os *transformers*, é o AdamW (LOSHCHILOV; HUTTER, 2019). O AdamW é uma versão aprimorada do algoritmo Adam, que se destaca por combinar a adaptação individual da taxa de aprendizado com a regularização L2, de forma desacoplada. Ele ajusta a taxa de aprendizado de cada parâmetro da rede com base no histórico de suas atualizações de gradiente, permitindo que parâmetros que mudam lentamente recebam atualizações maiores, enquanto aqueles que oscilam muito recebam ajustes menores. O principal diferencial do AdamW, como visto na Equação (2.11), reside na aplicação do *Weight decay* (uma forma de regularização L2) diretamente nos pesos do modelo, de forma separada das médias móveis dos gradientes (LOSHCHILOV; HUTTER, 2019). Essa abordagem desacoplada penaliza pesos muito grandes, promovendo soluções mais simples e generalizáveis, ao mesmo tempo em que evita as mudanças imprevisíveis que poderiam ocorrer se a regularização L2 fosse aplicada sobre as médias móveis.

Seja $g_t = \nabla_\theta \mathcal{L}(\theta_t)$ o gradiente da função de perda $\mathcal{L}$ em relação aos parâmetros θ no instante t. Assim como no Adam (ZHANG, 2018), são calculadas respectivamente as médias móveis de primeira e segunda ordem:

$$m_t = \beta_1 m_{t-1} + (1-\beta_1) g_t, \tag{2.8}$$

$$v_t = \beta_2 v_{t-1} + (1-\beta_2) g_t^2, \tag{2.9}$$

onde m_t corresponde à estimativa de primeira ordem (média dos gradientes) e v_t à estimativa de segunda ordem (média dos quadrados dos gradientes).

Para corrigir o viés inicial, utilizam-se as estimativas ajustadas:

$$\hat{m}_t = \frac{m_t}{1-\beta_1^t}, \quad \hat{v}_t = \frac{v_t}{1-\beta_2^t}. \tag{2.10}$$

A atualização dos parâmetros no AdamW é dada por:

$$\theta_{t+1} = \theta_t - \alpha \left(\frac{\hat{m}_t}{\sqrt{\hat{v}_t} + \epsilon} + \lambda \theta_t \right), \tag{2.11}$$

onde:

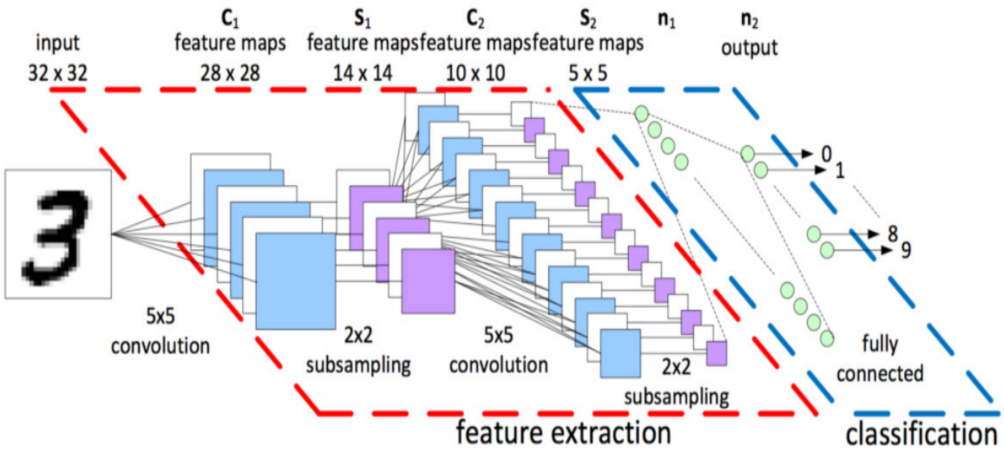
- α é a taxa de aprendizado,
- $\beta_1, \beta_2 \in [0, 1)$ são fatores de decaimento exponencial para os momentos de primeira e segunda ordem,
- ϵ é um termo de estabilidade numérica,
- λ é o coeficiente de *weight decay*.

2.1.4 Redes Neurais Convolucionais

As Redes Neurais Convolucionais, ou em inglês *Convolutional Neural Network* (CNN), são uma família das redes neurais tradicionais para imagens, que utilizam agora informações de uma vizinhança de *pixels* e lidam melhor com altas dimensionalidades, ao invés de apenas tratar a imagem como um simples vetor de números (ZHANG et al., 2023). Elas preservam a estrutura espacial dos dados ao aplicar convoluções — operações que capturam padrões locais, como bordas, texturas e formas. Tal característica é extremamente desejável para tarefas que envolvem diferenças estruturais sutis na natureza do problema, como a identificação do glaucoma. Essas redes utilizam camadas convolucionais, como mostrado na Figura 2.3, para extrair características hierárquicas, começando por elementos simples nas camadas iniciais e avançando para padrões mais complexos nas camadas mais profundas (ZHANG et al., 2023). Dado posto, se tornou muito popular em tarefas de classificação e reconhecimento de objetos em diversas áreas, como esporte (MARTIN et al., 2018), reconhecimento de emoções (KOSSAIFI et al., 2020), segurança (POOJARY; RAINA; KRISHANMURTHY, 2022), etc.

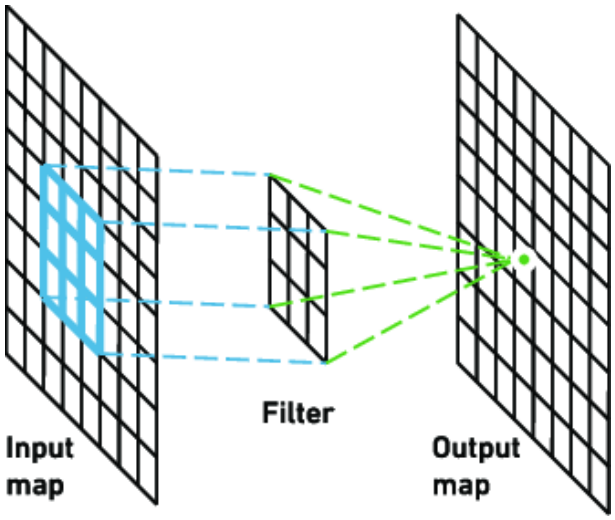
Aprofundando um pouco mais sobre as camadas convolucionais, elas são compostas basicamente pela aplicação de diversos filtros (também chamados de *kernels*) "deslizados" ao longo da imagem, assim extraindo características relevantes. A aplicação destes filtros se dá simplesmente pela aplicação de uma operação de multiplicação dos valores de uma matriz (o *kernel*) por cada um dos valores dos pixels correspondentes, como mostrado na Figura 2.4. Após isso, normalmente soma-se e toma-se a média dos valores para gerar um único valor que captou informações de toda uma vizinhança, explicando assim o potencial de caracterização dessa camada (ZHANG et al., 2023). A matriz resultante é denominada mapa de características.

Figura 2.3 – Exemplo de uma arquitetura CNN



Fonte: (BHATT et al., 2021)

Figura 2.4 – Exemplo de uma operação de convolução.



Fonte: (YAKURA et al., 2018)

Para manipular com maestria as redes, é importante conhecer as chamadas funções de ativação, que desempenham um papel fundamental nas CNNs, sendo responsáveis por introduzir não linearidade aos modelos, recebendo a entrada x do neurônio e aplicando tais transformações na mesma. Essa característica permite que a rede aprenda padrões complexos e sutis presentes nos dados, como os detalhes estruturais em imagens de fundo de olho associados ao glaucoma. Entre as funções de ativação mais utilizadas estão a *Rectified Linear Unit* (ReLU) (de simples implementação e compatível com quase todos os casos, resultando na sua escolha), que ativa apenas valores positivos e é eficiente em evitar o problema de saturação de gradientes, e a *Leaky-ReLU*, que resolve o problema do "*dying* ReLU" ao permitir pequenas ativações para valores negativos. Outras funções, como a sigmoide e a tangente hiperbólica, também são relevantes, mas tendem a ser menos utilizadas em redes profundas devido a problemas como gradientes desvanecentes (que diminuem muito e acabam prejudicando o aprendizado) (RASAMOELINA; ADJAILIA; SINČÁK, 2020). Porém, a sigmoide se mostra presente normalmente nas camadas

finais, juntamente da *softmax*, sendo usadas para dar as probabilidades finais das redes (KYURKCHIEV; MARKOV, 2015). Enquanto a primeira é comumente utilizada em problemas binários, a segunda costuma aparecer em problemas multi-classe (GAO; PAVEL, 2018).

A função sigmoide é definida matematicamente como:

$$\sigma(x) = \frac{1}{1 + e^{-x}} \tag{2.12}$$

onde:

- x corresponde a entrada de cada neurônio da camada, que pode ser qualquer valor real, geralmente resultado da sama ponderada da entrada da rede, incluindo pesos e vieses.
- e é a constante matemática de Euler, utilizada para calcular a transformação exponencial da entrada. Sua presença garante que a função sigmoide tenha um comportamento suave e contínuo.
- e^{-x} aplica uma transformação exponencial ao valor de ativação do neurônio, comprimindo-o para o intervalo entre 0 e 1.

A função *softmax* é definida matematicamente como:

$$\sigma(\mathbf{z})_i = \frac{e^{z_i}}{\sum_{j=1}^{K} e^{z_j}} \tag{2.13}$$

onde:

- $\mathbf{z} = [z_1, z_2, \ldots, z_K]$ é o vetor de entrada (logits) de dimensão K, contendo os valores de ativação para cada classe. Cada elemento z_i pode ser qualquer valor real.
- e^{z_i} aplica a transformação exponencial ao elemento z_i, convertendo valores negativos para positivos e amplificando diferenças entre os valores de entrada. Essa transformação é essencial para a interpretação probabilística.
- $\sum_{j=1}^{K} e^{z_j}$ é o termo de normalização (soma das exponenciais de todos os elementos do vetor). Garante que a soma das saídas seja igual a 1, transformando os valores em uma distribuição de probabilidade válida.

Já a função ReLU compartilha do mesmo x e é definida da seguinte forma:

$$\text{ReLU}(x) = \max(0, x). \tag{2.14}$$

Por fim, a adaptação da ReLU, a LeakyReLU é definida por:

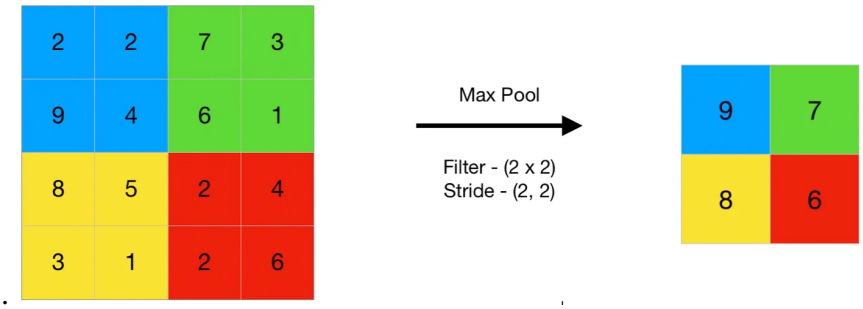
$$\text{LeakyReLU}(x) = \begin{cases} x, & \text{se } x \geq 0 \\ \alpha x, & \text{se } x < 0 \end{cases}, \tag{2.15}$$

onde:

- x é a entrada do neurônio, como anteriormente.
- α é um pequeno valor positivo (geralmente entre 0 e 1) que permite que a função *LeakyReLU* tenha uma pequena ativação para valores negativos, ajudando a evitar o problema do "*dying ReLU*", onde os neurônios podem ficar inativos e não aprender nada.

Outra característica das CNNs é a eficiência computacional, tendo em vista a facilidade de paralelização das operações de convolução em GPUs (ZHANG et al., 2023). As CNNs também comumente utilizam, além das convolucionais, algumas camadas chamadas de *pooling* para reduzir as dimensões espaciais dos mapas de características, mantendo as informações de maior magnitude. Essas camadas ajudam a controlar o número de parâmetros e, ao mesmo tempo, aumentam a capacidade de generalização do modelo, sendo as operações mais comuns as de *Max Pooling*, exemplificadas na Figura 2.5, e *Average Pooling* (GHOLAMALINEZHAD; KHOSRAVI, 2020).

Figura 2.5 – Exemplo de uma operação de *max pooling* utilizando um filtro 2×2 e um passo de *stride* de 2



Fonte: (KUMAR, 2023)

Na camada de *Max Pooling*, para cada região (normalmente uma janela de 2×2 ou 3×3) nos mapas de características, apenas o valor máximo é retido. Isso significa que, para cada região, apenas a informação mais proeminente (o valor máximo) é passada para a camada seguinte. A operação de *Max Pooling* ajuda a preservar características importantes e a reduzir o efeito de pequenas variações. Devido a essa característica, costuma ser mais comumente usado e, por isso, será utilizado nos experimentos. Já no *Average Pooling*, para cada região, a média dos valores é calculada e usada como representação. Isso suaviza a informação e ajuda a lidar com variações menores, tornando o modelo mais robusto a pequenas mudanças (GHOLAMALINEZHAD; KHOSRAVI, 2020).

2.1.5 *Transformers*

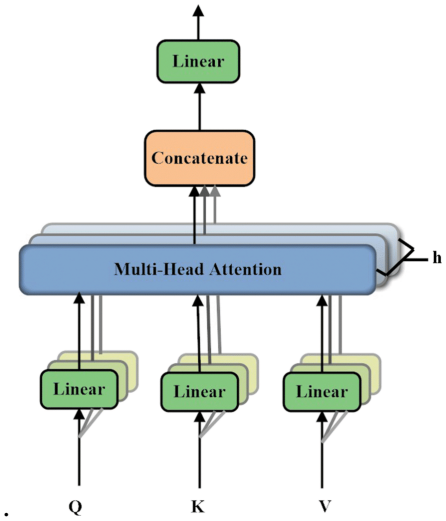
Outra solução cada vez mais popular dentre a literatura no campo médico, como visto em (TOHYE et al., 2024; DOSOVITSKIY et al., 2021), é a aplicação de *Transformers*. Estas redes foram introduzidas por Vaswani et al. (2017), substituindo as camadas de convolução por blocos baseados em atenção. O mecanismo central, chamado de *self-attention*, permite que cada

elemento da entrada (*token* ou fragmento de imagem) interaja diretamente com qualquer outro, capturando relações mais distantes de forma paralela e eficiente (BRAUWERS; FRASINCAR, 2023). Os *Transformers* se destacaram pela escalabilidade e velocidade de treinamento, o que levou a conquistas expressivas em tradução automática e modelagem de linguagem natural, que foi sua aplicação primária e pioneira (VASWANI et al., 2017).

Desde então, eles se consolidaram além do processamento de linguagem natural, abrangendo áudio, sinal, multimodal e visão computacional (ISLAM et al., 2024). Esses modelos diversas vezes superam redes neurais e CNNs tradicionais ao lidar com dependências de longa distância e processar dados em paralelo, embora exijam maior capacidade computacional. A evolução do paradigma gerou numerosas variantes como *Efficient Transformer*, *Longformer* e *Performer*, voltadas para reduzir custo computacional e memória sem perder precisão (TAY et al., 2022).

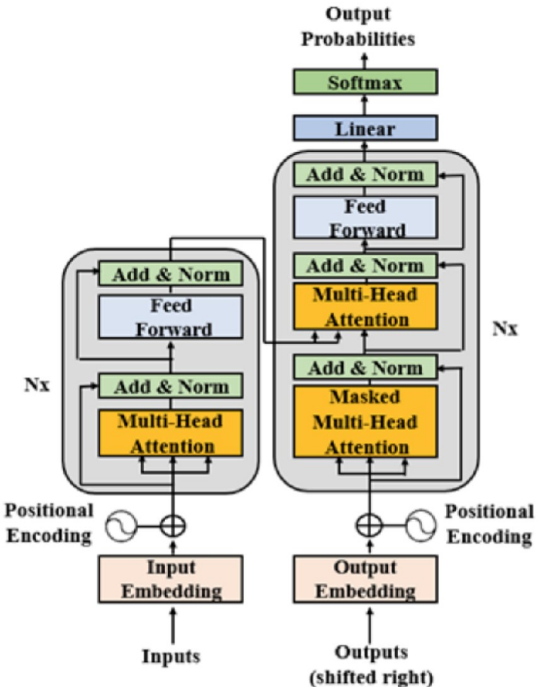
Na prática, os *Transformers* oferecem vantagens como atenção abrangente, paralelização e flexibilidade arquitetural. O uso de múltiplas cabeças de atenção (chamadas *multi-head attention* e vistas na Figura 2.6) e de codificações posicionais resolve a falta de ordenação temporal, permitindo o processamento eficaz de dados sequenciais ou visuais. A maneira como esses elementos atuam em conjunto pode ser observada na Figura 2.7. O sucesso contínuo em tarefas visuais levou ao desenvolvimento do ViT, que aplica o mesmo princípio de *tokens* a partes de imagens (DOSOVITSKIY et al., 2020).

Figura 2.6 – Exemplo de um *multi-head attention*, onde o Q, K e V representam vetores obtidos de cada *patch* por meio de multiplicações com matrizes de pesos aprendíveis



Fonte: (LUWEI et al., 2022)

Figura 2.7 – Exemplo da arquitetura básica de um *transformer*



Fonte: (SILVA, 2022)

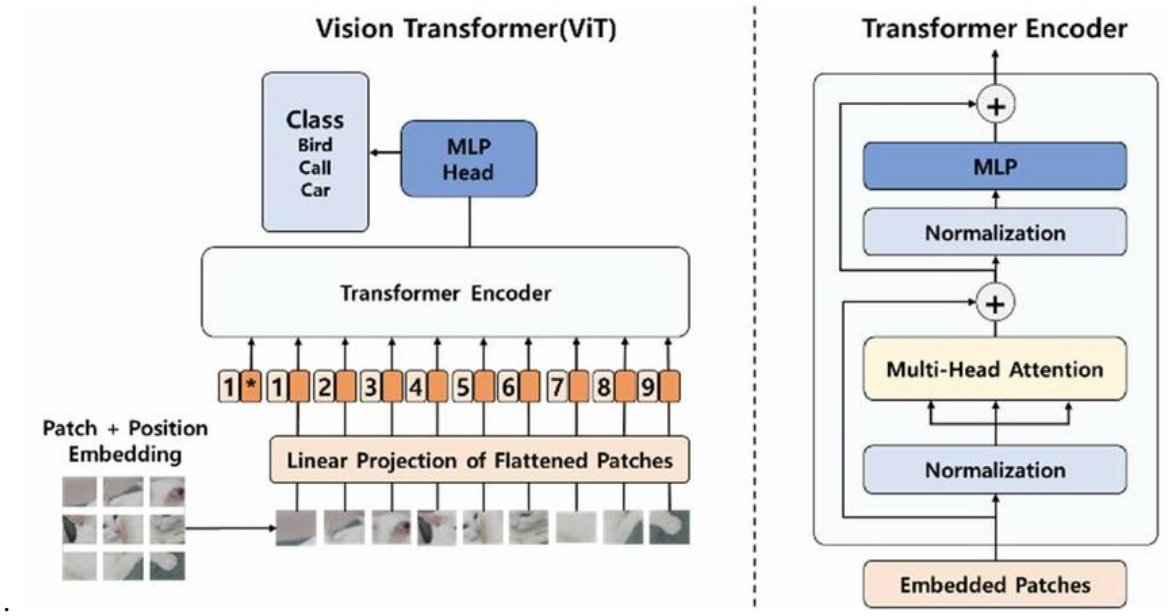
2.1.5.1 *Vision Transformer*

A arquitetura ViT, proposta por Dosovitskiy et al. (2020), adapta o *transformer* puro para visão computacional, tratando *patches* (pequenos pedaços) de imagem como *tokens*. Primeiramente, a imagem é dividida nestes *patches* com um tamanho fixo (como 16×16 *pixels*), cada um achatado e projetado linearmente em um espaço vetorial de dimensão D, formando os chamados *embeddings*. Incluem-se, então, *embeddings* posicionais aprendíveis e um *token* especial de classificação, chamado *Classification Token* (CLS), no início da sequência, convertendo a entrada em uma estrutura sequencial compatível com o *encoder Transformer* (DOSOVITSKIY et al., 2020). Um exemplo dela pode ser visto na Figura 2.8.

O *encoder* do ViT consiste em L camadas idênticas, cada uma composta por mecanismos de atenção chamados *Multi-Head Self-Attention*, seguidas de uma rede totalmente conectada, também chamada de *Multi-Layer Perceptron* (MLP), composta por uma ou mais camadas lineares intercaladas com funções de ativação não lineares, responsável por refinar as representações extraídas da atenção, juntamente com normalizações e conexões residuais (AWOFESO, 2024). Este mecanismo de atenção possibilita que cada *patch* "veja" todos os demais, modelando relações de longo alcance em toda a imagem. As conexões residuais e a normalização garantem estabilidade no treinamento, facilitando o fluxo de gradiente. Ao final, o vetor correspondente ao *token* CLS é extraído e alimenta o cabeçalho MLP simples para realizar a predição em si da classe da imagem (FACE, 2021).

Essa estrutura apresenta vantagens fundamentais: atenção global para modelar interde-

Figura 2.8 – Exemplo da arquitetura tanto do ViT à esquerda quanto do seu *encoder* à direita



Fonte: (BANG et al., 2023)

pendências visuais, paralelização eficiente por não depender de janelas ou recorrência, além de arquitetura modular, que facilita variações (como o MaxVit) se trocando pequenas partes dessa arquitetura (DOSOVITSKIY et al., 2020). No entanto, a obtenção de bons resultados requer pré-treinamento em grandes bases de dados (como JFT ou ImageNet-21k), pois o ViT, ao contrário das CNNs que assumem que *pixels* vizinhos são mais correlacionados, trata todas as partes igualmente, sem favorecer informação local, o que requer maiores volumes de dados para aprender relações espaciais (IDREES, 2024).

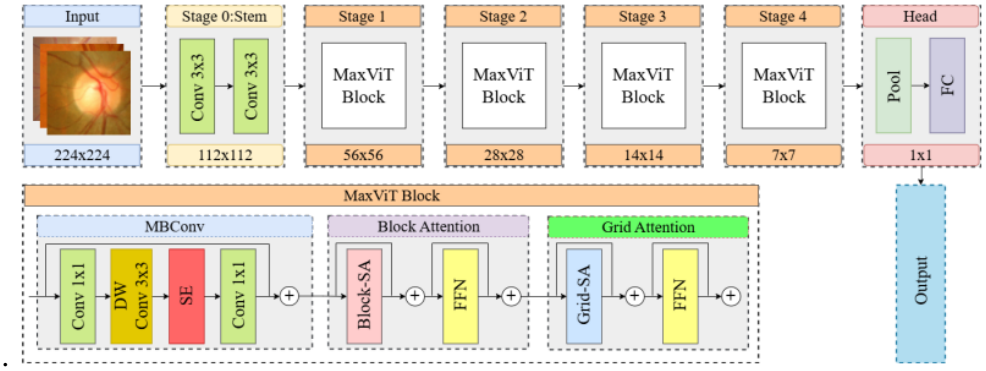
2.1.5.2 MaxViT

O MaxViT, introduzido por Tu et al. (2022), representa uma evolução do ViT, pois combina atenção local e global em uma única arquitetura híbrida que mescla convoluções (via MBConv) e mecanismos de *self-attention* em duas dimensões. Ao contrário do ViT original, que aplica atenção de forma global em todos os *patches* com complexidade quadrática, o MaxViT emprega um módulo chamado *Multi-Axis Self-Attention*, que primeiro aplica atenção por blocos em janelas locais e depois atenção por grade em regiões dilatadas, ambos com complexidade linear em relação ao tamanho da imagem. Essa abordagem permite que o modelo enxergue globalmente em estágios de alta resolução sem sobrecarregar o processamento, algo que o ViT puro não consegue (TU et al., 2022).

Como visto na Figura 2.9, a arquitetura é organizada em estágios, começando com um "*stem*" que utiliza convoluções 3×3 para reduzir a dimensão espacial e extrair características iniciais, seguida por repetidos blocos MaxViT. Cada bloco une MBConv a duas camadas de atenção (*block* e *grid attention*), intercaladas com normalizações e conexões residuais, garantindo treinamento estável e extração de padrões locais e globais. Ao fim da rede, a saída é reduzida

através de *pooling* e um cabeçalho completamente conectado gera a predição final (TU et al., 2022).

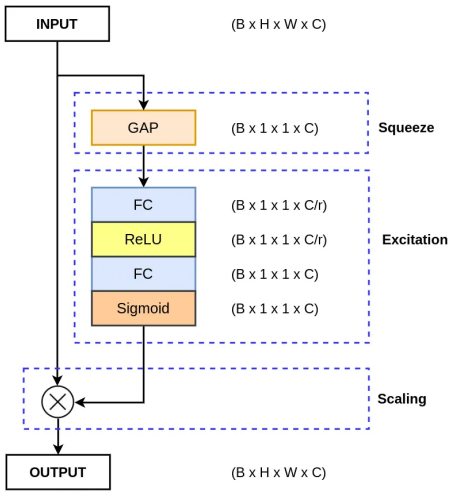
Figura 2.9 – Exemplo da arquitetura de um MaxViT



Fonte: (YURDAKUL; UYAR; TASDEMIR, 2025)

A MBConv utilizada implementa o chamado "*inverted bottleneck*", visto dentro do bloco na Figura 2.9, onde, primeiro, uma convolução 1×1 expande o número de canais. Em seguida, aplica-se uma convolução especial 3×3, que processa cada canal de forma independente, capturando padrões espaciais localmente para, finalmente, outra convolução 1×1 fazer a projeção de volta ao número original de canais. Além disso, esse bloco geralmente incorpora o módulo visto na Figura 2.10, chamado *Squeeze-and-Excitation*, que recalibra a importância dos canais com base no contexto global da imagem, melhorando assim a representação (YANG et al., 2022).

Figura 2.10 – Exemplo de um bloco *Squeeze-and-Excitation*, onde B, H, W e C representam respectivamente o *batch*, altura, largura e canal.



Fonte: (ERDOğAN, 2022)

Ainda que arquiteturas avançadas como o MaxViT tragam ganhos significativos em capacidade de representação e desempenho, seu sucesso na prática também depende de estratégias complementares que garantam estabilidade e generalização durante o treinamento, assim como

as CNNs e o ViT. Técnicas como regularização, normalização, aumento de dados e o uso de modelos pré-treinados no chamado *Transfer Learning* são fundamentais para mitigar problemas como o sobreajuste (ou *overfitting*), especialmente em domínios com dados limitados, como é o caso de aplicações médicas.

2.1.6 Normalização e regularização

No aprendizado profundo, fatores como a diversidade dos dados de entrada, a quantidade e a divisão dos dados, além do tempo de treinamento, podem levar o modelo a se ajustar excessivamente aos dados utilizados em sua construção. Esse fenômeno, conhecido como *overfitting* (ALPAYDIN, 2021), prejudica o desempenho do modelo ao ser aplicado em dados de teste. Em cenários como o proposto envolvendo imagens médicas em baixa quantidade, desbalanceados e semelhantes, esse problema se torna ainda mais presente. Para mitigá-lo, existem técnicas de regularização como o *Dropout*, bem como diferentes operações de normalização para facilitar o processamento.

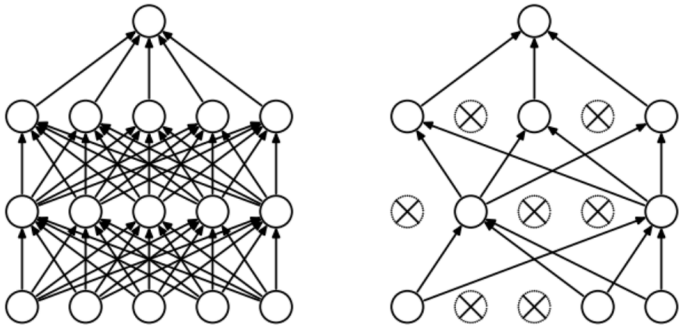
As técnicas de regularização são projetadas para restringir (ou aumentar) a complexidade dos modelos, também buscando evitar o *overfitting* e melhorar a capacidade de generalização. Entre as mais comuns estão a Regularização L1, a Regularização L2 e *Dropout*. A regularização L1 adiciona uma penalidade proporcional à soma dos valores absolutos dos pesos, promovendo a esparsividade no modelo. Já a regularização L2 adiciona uma penalidade proporcional à soma dos quadrados dos pesos, incentivando pesos menores e mais estáveis (MOORE; DENERO, 2011).

Já a regularização *Dropout* é uma técnica que consiste em, durante o treinamento, desativar, de forma aleatória, uma porcentagem dos neurônios em uma camada da rede neural. Essa desativação força o modelo a não depender excessivamente de determinados neurônios ou combinações específicas de características, promovendo uma maior generalização. No entanto, durante a fase de teste ou inferência, todos os neurônios permanecem ativos, e os pesos são ajustados proporcionalmente ao percentual de *Dropout* aplicado no treinamento (ZHANG et al., 2023).

As regularizações se mostram grandes aliadas no cenário de classificação proposto, visto que é uma tarefa extremamente complexa onde os dados tendem a ser semelhantes, com variações por vezes quase imperceptíveis. Tal fato acaba favorecendo bastante um possível *overfitting* do modelo, se adaptando demasiadamente às características gerais e não desejáveis presentes nas imagens.

Por fim, as técnicas de normalização são extremamente presentes no contexto das CNNs, porém dessa vez visando principalmente facilitar o processo de treinamento. Entre as técnicas mais comuns estão a normalização *min-max*, que reescala os valores dos *pixels* para uma faixa definida, geralmente entre 0 e 1, e a normalização *Z-score*, que ajusta os dados para que tenham

Figura 2.11 – Demonstração de uma aplicação de *dropout*. Em (a), está um exemplo de uma rede em seu estado normal, enquanto em (b) se tem a rede após aplicação da técnica.



Fonte: adaptado de(RAJ, 2023)

uma média de zero e um desvio padrão de um. Essas abordagens reduzem a influência de variações em iluminação, contraste, resolução, entre outros *outliers* que podem afetar o desempenho do modelo (SINGH; SINGH, 2019). Também vale o destaque para normalizações que aceleram a convergência do treinamento para um valor estável, como *Batch Normalization*, que reduz o deslocamento da covariância, assim garantindo estabilidade e agilidade, pois permite taxas de aprendizado maiores sem se preocupar com alguma variação brusca da distribuição dos dados a cada camada (BJORCK et al., 2018).

No contexto da classificação de imagens de fundo de olho apresentado no presente trabalho, uma normalização *min-max* se faz útil, visando minimizar algumas variações entre as imagens, como algumas mudanças sutis de iluminação, visto que o *dataset* utilizado é uma junção de diversos outros conjuntos. Além disso, a aceleração do treinamento do modelo é bem-vinda, prevenindo problemas com gradientes extremos, comumente presentes nos testes realizados.

2.1.7 *Data Augmentation*

Além das normalizações e regularizações, outra técnica amplamente utilizada em visão computacional visando a melhora dos modelos é o *data augmentation*, que consiste em criar dados sintéticos com diferentes variações, como rotações, translações, redimensionamentos, alterações de brilho, adição de ruído ou inversões de cores. Essas transformações geram novas instâncias que preservam as características relevantes das classes originais, atendendo a dois desafios principais enfrentados por pesquisadores na área: aumentar o volume de dados a partir de um conjunto limitado e reduzir o problema de *overfitting* (MUMUNI; MUMUNI, 2022).

Avanços significativos têm sido observados com a aplicação de técnicas de *data augmentation*, com diversas estratégias sendo desenvolvidas para expandir e melhorar a generalização de redes neurais. Por exemplo, o modelo AlexNet foi amplamente utilizado para avaliar e comparar diferentes métodos de aumento de dados nos conjuntos de dados ImageNet e CIFAR10, destacando-se a aplicação de rotações que apresentaram resultados superiores (MAHARANA;

MONDAL; NEMADE, 2022).

Um exemplo cada vez mais utilizado é o de redes que geram dados aumentados automaticamente durante o treinamento, reduzindo a perda do modelo. Existem até mesmo técnicas ainda mais avançadas, como *Generative Adversarial Networks* (GAN), que já foram aplicadas para criar imagens sintéticas de ressonância magnética com tumores cerebrais, por exemplo, destacando o potencial desse crescimento da quantidade de amostras para a melhora de desempenho em cenários de dados limitados (MAHARANA; MONDAL; NEMADE, 2022).

2.1.8 *Transfer Learning*

Por fim, finalizando as técnicas utilizadas comumente para melhoria das métricas de uma CNNs, dessa vez atuando na capacidade de captura de padrões e generalização nos dados, há o *Transfer Learning*. É uma técnica que também ajuda a reduzir a quantidade de dados necessários para treinar modelos eficientes, pois, ao invés de se treinar uma rede neural do zero, aproveitam-se pesos pré-treinados de um modelo desenvolvido para uma tarefa semelhante e os adapta a um novo problema (MORID; BORJALI; FIOL, 2021).

Essa abordagem é particularmente útil em domínios onde a disponibilidade de grandes volumes de dados rotulados é limitada, como na área médica, onde obter imagens anotadas pode ser um processo caro e demorado (KIM et al., 2022). Ao reutilizar camadas previamente treinadas em grandes bases de dados, como o *ImageNet*, as redes conseguem aprender representações mais genéricas das características das imagens, como formas, cores e objetos comuns, melhorando também a generalização final (MUREL; KAVLAKOGLU, 2025).

A adaptação dos modelos famosos como ResNet (KUNDU, 2023), VGG (SIMONYAN; ZISSERMAN, 2014) e EfficientNet (TAN; LE, 2019) pode ocorrer por meio de diferentes estratégias, como extração de características e ajuste fino. Na extração de características, utiliza-se a parte convolucional da rede pré-treinada como um extrator de atributos, adicionando novas camadas densas específicas para o contexto desejado (MORID; BORJALI; FIOL, 2021). Já no ajuste fino, além das novas camadas adicionadas, parte dos pesos das camadas pré-treinadas é ajustada durante o treinamento para melhor adaptação ao novo conjunto de dados (MUREL; KAVLAKOGLU, 2025).

2.1.9 Algoritmos de votação

Com o intuito de combinar as vantagens individuais de diferentes arquiteturas e mitigar erros que modelos singulares poderiam cometer, outra forma de utilização de modelos presente na literatura em aprendizado de máquina são as chamadas estratégias de votação, também conhecidas como métodos de *ensemble*. A ideia central é integrar as predições de múltiplos modelos para aumentar a robustez e a precisão do sistema preditivo (Scikit-learn, 2025).

A principal hipótese subjacente a esses métodos é que modelos com arquiteturas distintas

tendem a cometer erros em padrões diferentes. Por exemplo, CNNs são tipicamente mais sensíveis a características locais das imagens, como texturas e bordas, enquanto os modelos *Transformer* são mais eficazes em capturar dependências globais e relações de longo alcance. Ao combinar as previsões dessas abordagens complementares, é possível aproveitar as forças de cada uma e gerar um diagnóstico mais confiável. Trabalhos anteriores, como o de Velpula e Sharma (2023), já demonstraram a aplicação de sistemas de votação majoritária entre redes convolucionais para classificação de glaucoma.

No contexto de classificação, os esquemas de votação podem ser implementados de diferentes maneiras, considerando a natureza das saídas dos modelos:

- **Votação por Média (*Soft Voting*)**: Este esquema calcula a média das probabilidades preditas por cada modelo para cada classe. A classe final é determinada com base em um limiar de decisão, geralmente 0,5, para converter a probabilidade média em uma classificação binária. O *Soft Voting* utiliza a riqueza da informação probabilística fornecida pelos modelos individuais, sendo mais sensível e tendendo a funcionar melhor quando os modelos são bem calibrados. No entanto, sua performance pode ser negativamente impactada se um dos modelos participantes estiver descalibrado ou tiver um desempenho significativamente inferior, puxando a média para baixo (GeeksforGeeks, 2025).

- **Votação por Maioria (*Hard Voting*)**: Neste esquema, as probabilidades preditas por cada modelo são primeiramente convertidas em predições binárias (0 ou 1) usando um limiar de 0,5. Em seguida, a classe final é selecionada com base na maioria dos votos entre os modelos. Se houver um empate, uma regra de desempate predefinida é aplicada; no presente trabalho, a classe negativa foi considerada em caso de empate, evitando assim um número muito alto de diagnósticos errôneos, o que tornaria a classificação quase em um "chute". O *Hard Voting* é robusto a *outliers*, pois um modelo com desempenho muito inferior ou descalibrado contribuirá com apenas um voto para a decisão final, minimizando seu impacto negativo. Contudo, uma desvantagem é que ele ignora o grau de confiança de cada modelo, tratando uma probabilidade de 0,51 e uma de 0,99 da mesma forma, o que pode descartar informações valiosas (SIMIC, 2025).

A escolha entre *Soft* e *Hard Voting* depende das características dos modelos individuais (como os graus de confiança e proporção de *False Negatives* (FNs) e *False Positives* (FPs)) e do problema em questão, sendo importante avaliar a efetividade de cada abordagem na prática.

2.1.10 Métricas de avaliação de resultados

Para avaliar os resultados dos modelos de *machine learning*, são essenciais a utilização de métricas confiáveis que possam dar uma noção da eficácia final, especialmente em aplicações críticas como a detecção de glaucoma em imagens de fundo de olho. A correta escolha dessas

métricas permite avaliar não apenas a capacidade do modelo em identificar corretamente os casos de glaucoma, mas também em evitar falsos negativos, analisando métricas como o *recall* (ou revocação), que podem ter graves consequências clínicas. Entre as principais métricas utilizadas para problemas de classificação, temos a precisão, *F1-Score* e o citado *recall*.

A precisão mede a quantidade de vezes que o modelo acerta em relação ao total de vezes que ele tenta acertar. Enquanto isso, o *recall* mede a quantidade de vezes que o modelo acerta em relação ao total de vezes que ele deveria ter acertado. Por fim, o *F1-Score* é uma métrica que combina precisão e *recall* de maneira equilibrada, utilizando uma média harmônica entre ambas as métricas, garantindo assim que nenhuma seja demasiadamente discrepante da outra (FILHO, 2023).

Tabela 2.1 – Matriz de Confusão, onde as colunas representam as predições, e as linhas representam os rótulos reais

	Positive (P)	**Negative (N)**
Positive (P)	True Positive (TP)	False Negative (FN)
Negative (N)	False Positive (FP)	True Negative (TN)

Fonte: próprio autor.

Para entender como é feito o cálculo de cada uma destas métricas, é possível observar na Tabela 2.1 o que representam os conceitos de Verdadeiro Positivo, ou em inglês *True Positive* (TP), Falso Negativo, ou em inglês *False Negative* (FN), Falso Positivo, ou em inglês *False Positive* (FP), e Negativo Verdadeiro, ou em inglês *True Negative* (TN). Os TP são aquelas amostras corretamente classificadas como positivas pelo algoritmo, os TN são as instâncias corretamente classificadas como negativas pelo modelo, os FP são aquelas erroneamente classificadas como positivas pelo modelo e os FN são aquelas erroneamente classificadas como negativas pelo modelo (HOSSIN; SULAIMAN, 2015). Tendo isto em mente, as métricas utilizadas neste trabalho podem ser definidas da seguinte forma:

$$\text{Precision} = \frac{TP}{TP + FP}, \tag{2.16}$$

$$\text{Recall} = \frac{TP}{TP + FN}, \tag{2.17}$$

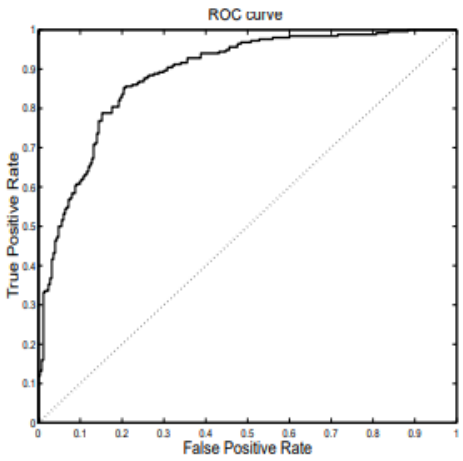
$$\text{F1-Score} = 2 \times \frac{\text{Precision} \times \text{Recall}}{\text{Precision} + \text{Recall}}. \tag{2.18}$$

Posto isto, ainda há outra métrica comumente utilizada como parâmetro de comparação na literatura: a acurácia. Ela indica a proporção de previsões corretas feitas pelo modelo em relação ao total de previsões. Sua popularidade se dá à sua simplicidade e facilidade de interpretação, sendo adequada para problemas de classificação onde as classes estão bem balanceadas. No entanto, em cenários com desequilíbrio significativo entre as classes como o presente, a acurácia

pode ser enganosa, pois um modelo pode apresentar alta acurácia ao prever corretamente a classe majoritária, enquanto falha em detectar a classe minoritária (HOSSIN; SULAIMAN, 2015). Devido a este fato, é necessário levá-la em consideração em conjunto com outras métricas, como o *recall* que se encaixa no contexto de saúde pelo fato de que se quer recuperar o máximo possível de previsões corretas.

Por fim, devido a esse fator, a literatura na área também costuma utilizar o conceito de AUC, que é uma métrica que avalia a capacidade de um modelo de classificação em distinguir entre classes. Ela mede a área sob a curva *Receiver Operating Characteristic* (ROC), que é uma curva que relaciona o *recall* com a taxa de falsos positivos em diferentes limiares de decisão, exemplificado na Figura 2.12. Uma AUC de 1 indica um modelo perfeito, enquanto uma AUC de 0,5 indica desempenho equivalente a uma classificação aleatória.

Figura 2.12 – Exemplo de uma curva *Receiver Operating Characteristic*.



Fonte: (RAKOTOMAMONJY, 2004)

2.2 Trabalhos Relacionados

A classificação de glaucoma em imagens de fundo de olho, embora esteja em rápida evolução e expansão, ainda enfrenta diversos desafios. Grande parte desses obstáculos decorre da baixa sensibilidade às mudanças estruturais visualmente perceptíveis, além da heterogeneidade das imagens da comorbidade em estágios iniciais (RODRIGUES et al., 2016). Para mitigar essas limitações, a integração de sistemas inteligentes de visão computacional tem se mostrado de grande valia no diagnóstico e classificação. Entre as abordagens mais recorrentes na literatura estão as Redes Neurais Convolucionais, Redes com Autoencoders, Redes de Atenção e metodologias híbridas que combinam essas técnicas. Contudo, um dos principais desafios permanece na obtenção de dados, devido à escassez de bases de dados consideráveis e de alta qualidade disponíveis para pesquisas de acesso livre (ASHTARI-MAJLAN; DEHSHIBI; MASIP, 2023).

Destacando a relevância da qualidade dos dados nessa área, o autor Li et al. (2018) concentrou seu esforço em garantir a consistência dos dados antes de aplicar qualquer modelo de

aprendizado de máquina. Para isso, utilizou cerca de 20 oftalmologistas na seleção, classificação e rotulação de imagens provenientes de diversos hospitais na China, empregando um sistema de consenso entre três especialistas para a tomada de decisão em cada amostra. Com os dados assim refinados, foi utilizada uma Inception-V3 simples, que alcançou uma AUC próxima de 0,99, apresentando dificuldades apenas em imagens com múltiplas doenças. No entanto, replicar essa abordagem em outras regiões, especialmente em áreas subdesenvolvidas como o Brasil, pode ser desafiador devido à disponibilidade limitada de especialistas e ao tempo necessário para tais análises.

Ao adotar uma análise mais generalista, trabalhos como os de Serte e Serener (2019) e Zhen et al. (2018) investigam o desempenho bruto de algumas das principais redes pré-treinadas conhecidas na literatura. Em (ZHEN et al., 2018), são analisadas redes como VGG-16, VGG-19, ResNet, DenseNet, InceptionV3, InceptionResNet, Xception e NASNetMobile, enquanto em (SERTE; SERENER, 2019) as abordagens se concentram na ResNet-50, GoogLeNet e ResNet-152, utilizando tanto pesos pré-treinados quanto as arquiteturas não ajustadas. Ambos os estudos aplicaram técnicas de segmentação do disco óptico — uma região frequentemente associada a diferenças indicativas da doença — e estratégias de *data augmentation* clássicas para minimizar o *overfitting* dos modelos. Além disso, em (SERTE; SERENER, 2019), foi realizada uma abordagem com múltiplas bases públicas (*High-Resolution Fundus* (HRF), Drishti-GS1, RIM-ONE, sjchoi86-HRF, ACRIMA), alternando os conjuntos de treino e teste por meio de permutações entre essas bases. Apesar das estratégias adotadas, nenhum dos estudos superou 80% de acurácia, o que evidenciou que o uso direto dos pesos da ImageNet não é suficiente para capturar informações discriminantes nesse contexto. Isso ressalta que a tarefa não é suficientemente generalista para os modelos amplamente utilizados atualmente, exigindo esforços adicionais para melhorar a performance.

No estudo de U et al. (2018), os autores propuseram uma abordagem básica fundamentada na eliminação da necessidade de pré-processamentos custosos, uma prática amplamente adotada nas pesquisas da área. Foi desenvolvida uma CNN com 18 camadas, estruturada com camadas convolucionais, camadas de *pooling* e normalização por *batch*. O modelo foi treinado com 1.426 imagens completas de fundo de olho, apenas redimensionadas, provenientes de uma base privada de um hospital na Índia. Apesar da simplicidade, o modelo atingiu cerca de 98% de acurácia, mostrando-se uma solução prática e de fácil integração devido à ausência de modificações prévias complexas. No entanto, devido à especificidade dos dados utilizados, não é possível garantir que o desempenho seja reproduzido em contextos mais gerais.

Quando observadas abordagens semelhantes porém mais compactas em tamanho de rede, têm-se (SAXENA et al., 2020), onde foi proposta uma rede convolucional profunda com seis camadas: quatro camadas convolucionais e duas camadas totalmente conectadas. Como entrada era passada uma imagem da região de interesse, ou *Region of Interest* (ROI), extraída através da divisão da mesma em grade para identificação do nervo óptico, redimensionada e

equalizada em iluminação. Porém, não foram fornecidos detalhes específicos desse processo de segmentação, dificultando reproduções seguintes. A estratégia de aumento de dados escolhida foi uma média das previsões realizadas pela camada *soft-max* da rede em recortes realizados em cada amostra, levando a resultados de 0,882 e 0,822 para o AUC nos datasets SCES e *Online Retinal Fundus Image Dataset for Glaucoma Analysis and Research* (ORIGA) respectivamente, que são bases extremamente utilizadas na área. Embora o artigo tenha dado uma visão bem fundamental, ele não apresentou nenhuma abordagem inovadora ou avanço metodológico. Como resultado, os resultados obtidos refletem essa limitação, permanecendo no nível básico e sem contribuir significativamente para o avanço da área.

Já em (CHEN et al., 2015), os autores exploraram o uso de múltiplas CNNs autorais, cada uma composta por seis camadas. Nessa proposta, as redes foram conectadas sequencialmente, de modo que a saída de uma rede ***i*** era utilizada como contexto para a última camada da rede subsequente ***i+1***. Para melhorar a robustez do modelo, foi adotada uma estratégia de aumento de dados baseada em recortes específicos, incluindo os cantos e o centro das imagens, que foram processados com técnicas como *flips* e redimensionamentos. A previsão final de cada amostra foi obtida por meio da média das previsões realizadas para todos os recortes. Combinando essa abordagem com a segmentação do disco óptico, previamente utilizada em outros trabalhos, os autores alcançaram uma AUC de 0,838 no conjunto de dados ORIGA e 0,898 no conjunto SCES, utilizando uma configuração de cinco redes conectadas conforme descrito. Apesar dos resultados promissores, é necessário ponderar o elevado custo computacional envolvido no treinamento simultâneo de cinco redes, mesmo que elas não sejam excessivamente complexas.

No estudo de Wu et al. (2020), foi proposta uma abordagem baseada no paradigma *teacher-student*, em que uma rede transfere conhecimento para outra, com o objetivo de captar padrões que poderiam ser menos perceptíveis para uma única rede. O modelo *teacher* utiliza imagens combinadas com máscaras de segmentação do disco óptico, mas sem a classificação quanto à presença do glaucoma. Por sua vez, o modelo *student* trabalha com imagens de treinamento rotuladas para indicar a presença ou ausência da doença. Durante o processo, ocorre a destilação de conhecimento, e, após uma rodada, o *student* é avaliado com dados não rotulados (denominados *quiz pool*). Os resultados dessa avaliação são então utilizados para atualizar o *teacher* e após todo o processo, foi alcançado cerca de 95% de acurácia na base LAG. Apesar de permitir que a rede principal se torne robusta sem a necessidade de ser muito complexa, essa abordagem apresenta elevada complexidade computacional durante o treinamento.

Outra abordagem baseada na utilização de múltiplas CNNs foi apresentada por Velpula e Sharma (2023), onde diversas redes pré-treinadas distintas foram empregadas, e uma fusão de seus resultados foi realizada por meio de uma abordagem de votação majoritária, ou seja, baseada na escolha da maioria. As redes utilizadas incluem ResNet50, AlexNet, VGG19, DenseNet-201 e Inception-ResNet-v2, aplicadas à classificação de imagens de quatro bases de dados distintas: ACRIMA, HVD, RIM-ONE e Drishti. Por meio dessa estratégia simples de fusão, a

implementação alcançou resultados com elevada acurácia, com valores variando de 99,57% na base ACRIMA até 85,43% na HVD. No entanto, esses resultados destacam, mais uma vez, a possível dificuldade de generalização dos modelos experimentais da área, devido à variedade dos dados. Além disso, questiona-se a viabilidade prática dessa abordagem em termos de tempo de execução, pois a utilização de múltiplos modelos aumenta o tempo de predição em até cinco vezes em comparação com o desempenho de um único modelo.

Finalizando os trabalhos unicamente baseados em CNNs, o estudo de Velpula et al. (2024), tem uma proposta baseada na combinação de redes neurais convolucionais e classificadores de aprendizado de máquina clássicos como o *Support Vector Machine* (SVM). A metodologia emprega redes pré-treinadas como a *ResNet18* e *ResNet50*, para a extração de características das imagens, capazes de capturar padrões complexos associados ao glaucoma, que são subsequentemente utilizadas por classificadores como o SVM e *K-Nearest Neighbors* (KNN), para distinguir entre imagens de olhos saudáveis ou não. Os resultados alcançados no conjunto de dados ACRIMA demonstraram uma acurácia de 98,03% ao utilizar o modelo ResNet50 para a extração de características e o classificador SVM. Apesar do resultado alto, por ter sido testado em uma base pequena e muito específica, é necessária uma análise mais aprofundada sobre a generalização do modelo para dados mais diversificados.

Buscando abordar a lacuna na comparação entre diferentes arquiteturas em contextos clínicos, saindo apenas do olhar para CNNs, Hwang et al. (2023) conduziu um estudo *multi-dataset* para comparar ViTs e CNNs, especificamente o modelo *ResNet-50*, na detecção do glaucoma. Para isso, seis bancos de dados públicos foram utilizados, incluindo Drishti-GS1, sjchoi86-HRF, RIM-ONE DL, ORIGA, ACRIMA e REFUGE2. Esses conjuntos de dados variam em tamanho (de 101 a 800 imagens) e apresentam diferentes graus de desequilíbrio entre as classes positivas e negativas. As imagens foram pré-processadas para extrair e recortar a região da cabeça do nervo óptico, usando segmentação semântica via deeplabv3plus. Ambos os modelos foram treinados em 80% dos dados e testados nos 20% restantes, com o ViT sendo pré-treinado no *ImageNet*. Os resultados indicaram que os modelos ViTs frequentemente superaram as CNNs em termos de AUC, acurácia e *F1-score* em cinco dos seis conjuntos de dados, especialmente naqueles com maior proporção de imagens da classe negativa. Por exemplo, no REFUGE2, o ViT alcançou um AUC de 0.95, enquanto o CNN obteve 0.89. Contudo, apesar da maior sensibilidade observada nos ViTs, eles tenderam a produzir mais falsos positivos em alguns conjuntos de dados, resultando em menor especificidade em certos cenários, como no ACRIMA. Esse achado sugere que, embora os modelos de *transformers* capturem relações globais de forma eficaz, sua maior dependência de uma representação de classe suficiente durante o treinamento pode levar a especificidades mais baixas em cenários de desequilíbrio de classes.

Trazendo esse movimento e saindo das técnicas mais difundidas baseadas em CNNs, no trabalho apresentado em (TOHYE et al., 2024), tem-se uma utilização mais aplicada de ViT. Nesse trabalho, foi empregada uma arquitetura apelidada de *Contour-Guided and Augmented*

Vision Transformer (CA-ViT), técnica essa que utiliza um *Conditional Variational Generative Adversarial Network* (CVGAN) para aumentar e diversificar o conjunto de dados de treinamento por meio da geração e reconstrução de amostras condicionais. Em seguida, é extraído o contorno tanto das imagens originais quanto das geradas especialmente nas regiões do disco óptico. As imagens originais, juntamente com seus contornos, são então enviadas ao *backbone* do ViT para treinamento, buscando tornar o modelo mais robusto a variações de cor e iluminação, além de direcionar a análise para áreas de maior relevância. Embora o uso de *transformers* seja relativamente recente no contexto de diagnóstico por imagens, o estudo obteve um resultado promissor, alcançando 93% de acurácia na SMDG-19. Esse resultado ressalta o potencial dessa abordagem e aponta para a necessidade de mais investigações no campo. Entretanto, vale destacar a grande dependência de ViTs por bases de dados extensas (DOSOVITSKIY et al., 2021), o que representa um desafio relevante, especialmente na área da saúde, devido à escassez de bases públicas de qualidade de tamanho adequado .

Também no campo dos trabalhos mais recentes com a utilização de *transformers*, já são observadas abordagens inovadoras que vão além da simples classificação. Em (CHINCHOLI; KOESTLER, 2024), o objetivo foi a detecção do *Optic Disc* (OD) e do *Optic Cup* (OC), utilizando modelos como ViTs e *Detection Transformer* (DETR). Em vez de se limitarem à classificação, esses modelos foram empregados para identificar as regiões do OD e do OC, permitindo a análise de características específicas indicativas de glaucoma, como a relação entre os diâmetros do OC e do OD. Essa proporção, geralmente considerada indicativa de glaucoma quando ultrapassa 0,6, é calculada a partir das detecções feitas pelos modelos. Os resultados mostraram-se promissores, com o DETR alcançando uma acurácia de 90,48% no conjunto de dados SMDG-19, também utilizado no presente trabalho, superando o ViT, que obteve 87,87%. No entanto, o treinamento dos modelos foi realizado em uma infraestrutura robusta e demandou cerca de 1.000 épocas, o que pode dificultar sua aplicação em contextos práticos.

Indo para variações dos *transformers*, no estudo de Yurdakul, Uyar e Tasdemir (2025), uma abordagem especializada para a classificação de glaucoma, chamada MaxGlaViT, foi proposta a partir do MaxViT, passando por três fases de aprimoramento: primeiro, o MaxViT foi escalonado para otimizar o número de blocos e canais, resultando em uma arquitetura mais leve (6.2 milhões de parâmetros contra os 31 milhões da versão mais leve do MaxVit, chamada de *Tiny*), buscando evitar overfitting. Em segundo lugar, o bloco *stem* do MaxViT foi melhorado com a adição de mecanismos de atenção após as camadas de convolução, permitindo que o modelo selecionasse características importantes e ignorasse informações irrelevantes. Por fim, as estruturas *MBConv* nos blocos MaxViT foram substituídas por blocos de *deep learning* avançados, atuando na melhoria da capacidade de generalização do modelo. O modelo foi avaliado no conjunto de dados HDV1, que inclui imagens de glaucoma avançado, inicial e normal, apresentando 92.03% de acurácia, 92.33% de precisão, 92.03% de recall e 92.13% de F1-score. Quando comparado com resultados da literatura utilizando o mesmo conjunto de dados, o modelo proposto aumentou a acurácia em 5.71%. Apesar do desempenho superior, o artigo reconhece que

desafios na generalização e interpretabilidade de soluções existentes persistem. Implicitamente, o fato de que “os modelos pequenos tiveram melhor desempenho no conjunto de dados”, dada a sua natureza pequena, com apenas 1542 imagens, pode sugerir uma dependência do tamanho otimizado para evitar *overfitting*, o que poderia levantar questões sobre sua robustez e necessidade de reescalonamento em datasets significativamente maiores e mais diversos.

A análise da literatura mostra que a área de classificação de glaucoma com base em imagens de fundo de olho é extremamente ampla, apresentando abordagens bastante distintas. No entanto, um fator em comum pode ser identificado: a grande variedade de bases de dados disponíveis, muitas delas com um número muito reduzido de imagens, o que compromete a comparação entre os estudos e a avaliação da capacidade de generalização dos modelos. Para abordar essa questão, o presente trabalho utilizará a SMDG-19, que combina 19 bases de dados públicas, proporcionando uma maior diversidade de amostras, conforme realizado por Tohye et al. (2024) e Chincholi e Koestler (2024).

3 Metodologia Proposta

Enquanto nos capítulos anteriores foram apresentados conceitos necessários para o entendimento da pesquisa, além de soluções já presentes no ramo da classificação de glaucoma, neste capítulo será apresentado o método proposto, visando trazer alguns métodos mais clássicos e difundidos para um conjunto de dados mais completo. Na Seção 3.1 é apresentada uma visão geral sobre esse conjunto, enquanto na Seção 3.2 são descritos os modelos em si, com detalhes sobre camadas e escolhas realizadas, juntamente do algoritmo utilizado para a predição. Por fim, na Seção 3.3 é descrita a forma como os resultados serão discutidos e medidos.

3.1 Base de Dados

Para esse estudo, foi utilizada a base *Standardized Multi-Channel Dataset for Glaucoma* (SMDG-19), que é uma base formada através da união de 19 *datasets* públicos comumente utilizados na literatura, podendo ser observada a relação disponível de cada um deles na Tabela 3.1. Devido a esse esforço, ela é o maior conjunto de dados públicos disponível, tanto para classificação quanto para segmentações específicas como de OD e até mesmo de vasos sanguíneos.

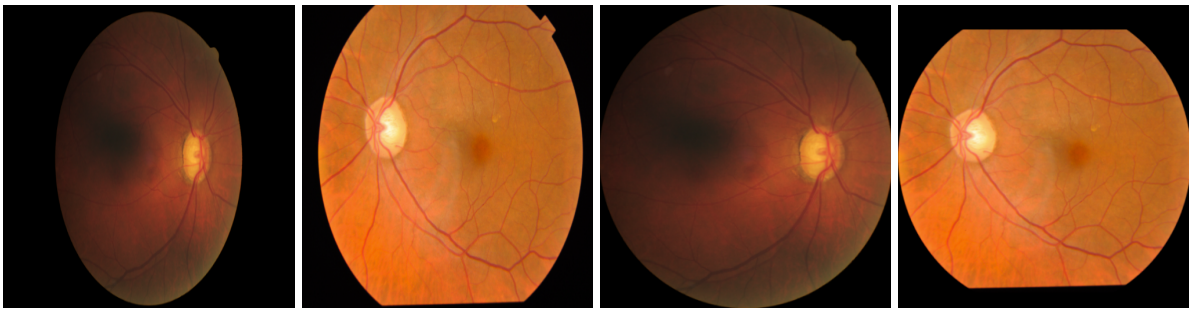
Tabela 3.1 – Resumo dos Conjuntos de Dados.

Conjunto de Dados	**Não-Glaucoma**	**Glaucoma**	**Suspeito**
BEH (Bangladesh Eye Hospital)	463	171	0
CRFO-v4	31	48	0
DR-HAGIS	0	10	0
DRISHTI-GS1-TRAIN	18	32	0
DRISHTI-GS1-TEST	13	38	0
EyePACS-AIROGS	0	3.269	0
FIVES	200	200	0
G1020	724	296	0
HRF (High Resolution Fundus)	15	15	0
JSIEC-1000	38	0	13
LES-AV	11	11	0
OIA-ODIR-TRAIN	2.932	197	18
OIA-ODIR-TEST-ONLINE	802	58	25
OIA-ODIR-TEST-OFFLINE	417	36	9
ORIGA-light	482	168	0
PAPILA	333	87	68
REFUGE1-TRAIN	360	40	0
REFUGE1-VALIDATION	360	40	0
sjchoi86-HRF	300	101	0
Total	**7.499**	**4.817**	**133**

Fonte: adaptado de (KIEFER, 2023).

As imagens da base foram padronizadas pelo autor Kiefer (2023), seguindo um algoritmo que recorta o fundo, centraliza a imagem do fundo de olho, preenche as informações ausentes e redimensiona para 512×512 *pixels*. Tal metodologia foi baseada em estudos anteriores realizados em (KIEFER et al., 2023), visando garantir a consistência e a qualidade das imagens para aplicações em aprendizado de máquina, afinal se tratando de uma junção de fontes distintas, é natural a existência de inconsistências que variam desde a posição até mesmo tamanho e cor. Um exemplo dos resultados destas mudanças pode ser observado na Figura 3.1.

Figura 3.1 – Exemplos de imagens após padronização. As imagens (a) e (b) representam as versões originais das figuras (c) e (d) respectivamente



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

Fonte: próprio autor.

O conjunto de dados está disponível no Kaggle [1], tornando o acesso live e visando fomentar o estudo de glaucoma com aprendizado de máquina, com a promessa de atualizações futuras à medida que mais conjuntos públicos forem publicados. Além disso, ele conta também com um dicionário de metadados contendo diversas informações extras que podem ser utilizadas para outras análises (KIEFER, 2023).

Na literatura, já é possível encontrar trabalhos utilizando este *dataset*, mesmo se tratando de um conjunto recente, como (TOHYE et al., 2024) e (DOSOVITSKIY et al., 2021), porém, ambos utilizam abordagens distintas, utilizando geração de imagens artificiais e relação de diâmetro entre OC e OD. Por isso, um dos objetivos será justamente aplicar técnicas mais clássicas e distintas nesse conjunto, como na votação de forma semelhante à utilizada por Velpula e Sharma (2023).

3.1.1 Pré-processamento e divisão dos dados para treinamento

Uma etapa essencial do treinamento de modelos é todo o tratamento aplicado aos dados, garantindo assim um comportamento mais controlado, estável e menos propenso a erros como gradientes anormais. Como a metodologia engloba modelos bem distintos com CNNs e ViTs, cada um deles exige pré-processamentos distintos.

[1] Disponível em: <https://www.kaggle.com/datasets/deathtrooper/multichannel-glaucoma-benchmark-dataset/data>. Acesso em Fevereiro de 2025.

3.1.1.1 Pré-processamento para as CNNs

O primeiro passo é o redimensionamento das imagens de fundo de olho para o tamanho esperado pela entrada da rede de 124×124 *pixels*.

Passada essa etapa, as imagens já redimensionadas são normalizadas através da divisão dos valores de cada um de seus *pixels* por 255, garantindo assim que todos se encontrem no intervalo entre 0 e 1. Com isso, variações bruscas e ruídos são mitigados, além de garantir estabilidade numérica (SINGH; SINGH, 2019). Um ajuste também necessário é realizado em relação às classes e rótulos, devido à escolha pela não utilização da classe "Suspeito", originalmente presente nos dados, resultando na exclusão de todas essas amostras. A decisão foi motivada pela baixa representatividade da classe no conjunto, tendo pouco mais de 1% dos dados totais, podendo prejudicar métricas e também não contribuindo muito para o objetivo principal de identificação da patologia.

3.1.1.2 Pré-processamento para os modelos baseados em *Transformers*

No caso dos modelos baseados em *Transformers* (ViT e MaxViT), o pré-processamento também inicia com a leitura e filtragem das imagens válidas. Assim como na etapa anterior, são consideradas apenas imagens coloridas e rotuladas como saudáveis ou não, com a exclusão de amostras rotuladas como Suspeito, devido à baixa representatividade da classe nos dados, o que poderia comprometer a robustez das métricas.

As imagens válidas são então redimensionadas conforme as exigências dos modelos base utilizados: para o ViT, o redimensionamento é feito para 224×224 *pixels*, enquanto no MaxViT, o tamanho utilizado é de 384×384 *pixels*. Em ambos os casos, o redimensionamento é seguido da conversão para tensores e normalização dos valores dos *pixels*, utilizando médias e desvios padrão adequados. Especificamente, no MaxViT, é aplicada uma normalização com média $[0.5, 0.5, 0.5]$ e desvio padrão $[0.5, 0.5, 0.5]$ por canal, garantindo que os valores fiquem aproximadamente entre $[-1, 1]$, o que facilita a convergência do modelo. Já no ViT a normalização ocorre internamente pela biblioteca do *Hugging Face*.

3.1.1.3 Divisão de dados e aumento de dados

Para divisão dos dados, foi adotada uma divisão de 70% dos dados para treinamento, enquanto 15% vão para teste e outros 15% vão para validação. A escolha de tais valores foi baseada em alguns dos trabalhos recentes com melhor desempenho (TOHYE et al., 2024), possibilitando assim uma base comparativa direta. A distribuição das classes em cada partição foi monitorada para garantir balanceamento razoável, mantendo sempre algo próximo de 60% dos dados da classe saudável e 40% dos dados da classe não saudável.

Finalizando o pré-processamento, no caso das CNNs é aplicado um *Data Augmentation*, conforme indicado na Tabela 3.2, visando assim melhorar a generalização da rede com dados

artificialmente transformados dinamicamente durante o treinamento, através do módulo *ImageDataGenerator* e do método *flow*, do próprio *TensorFlow*. De maneira simples, esse módulo aplica aleatoriamente aos *batchs* as transformações definidas, formando dessa forma um conjunto de dados mais diverso.

Tabela 3.2 – Parâmetros utilizados para *Data Augmentation*.

Parâmetro	**Descrição**
horizontal_flip	Espelha a imagem horizontalmente
vertical_flip	Espelha a imagem verticalmente
rotation_range	Rotação aleatória de até 45 graus
height_shift_range	Deslocamento vertical de até 20%
width_shift_range	Deslocamento horizontal de até 20%
zoom_range	Zoom aleatório de até 20%
fill_mode	Método de preenchimento: nearest

Fonte: próprio autor.

Para os *transformers* não foi realizado nenhum nível de *data augmentation*. Estudos como (STEINER et al., 2022) demonstram que, quando os modelos são pré-treinados em grandes bases de dados (como as variantes da ImageNet), eles podem obter bom desempenho mesmo com aumento mínimo ou ausência de *data augmentation*, sobretudo se o ajuste fino for realizado em conjuntos médios ou pequenos. Como o trabalho em questão se enquadra nesse aspecto, foi escolhida a não utilização de tal técnica, até como forma de analisar de maneira mais pura o desempenho desses modelos mais complexos.

3.2 Metodologia Proposta

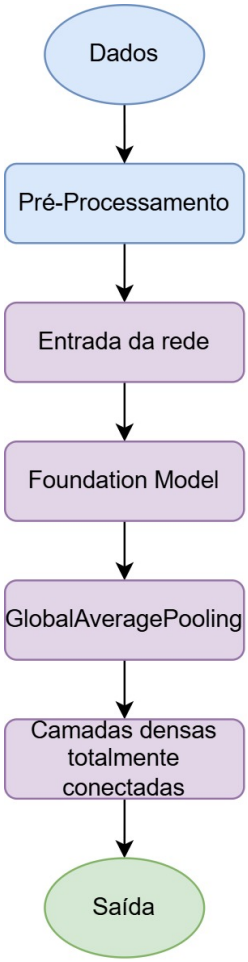
3.2.1 CNNs

De primeiro momento, para os modelos utilizando apenas redes convolucionais, foi estabelecida uma base comum de modelos, observada na Figura 3.2, visando-se ter um ponto comum para se avaliar diferenças entre diferentes modelos fundadores.

A rede começa com uma camada de entrada que redimensiona os dados para um tamanho de 124×124 *pixels* no espectro RGB, visando reduzir o custo computacional por limitações do hardware usado para os experimentos, mas tentando ao máximo preservar a qualidade para obter informações relevantes o suficiente para a tarefa.

Com os dados já devidamente redimensionados, a base convolucional do modelo se dá através de diferentes modelos fundadores, todos com os pesos inicializados da *Imagenet*, aproveitando-se de suas camadas convolucionais apenas e de sua identificação inicial de formas básicas. Nessa etapa, têm-se três variações de modelos a serem testadas: a ResNet-50, a VGG-19 e a EfficientNetB0, sendo brevemente explicadas a seguir.

Figura 3.2 – Arquitetura do modelo base, onde as partes em azul representam as etapas relacionadas ao tratamento dos dados, a roxa as etapas do modelo e de verde se encontra a saída final.



Fonte: próprio autor

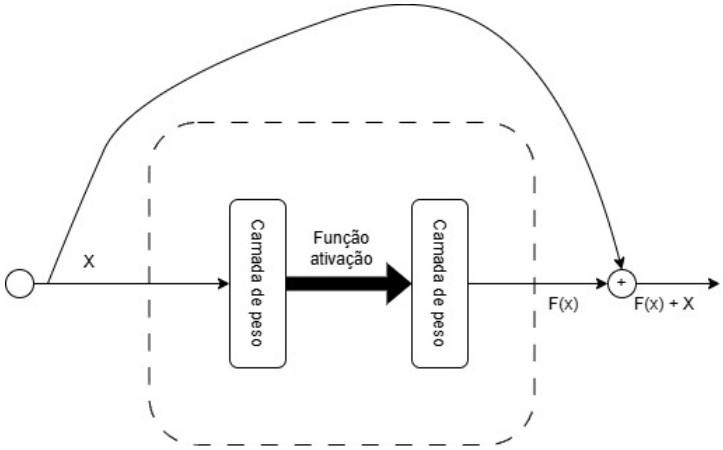
3.2.1.1 ResNet-50

A ResNet-50 é uma arquitetura de rede neural profunda composta por 50 camadas, desenvolvida para superar o problema do gradiente desaparecendo em redes muito complexas. Utilizando conexões residuais, como exemplificadas na Figura 3.3, gradientes anteriores fluem para camadas posteriores dentro de um bloco de convolução, evitando assim perdas em redes profundas como a adotada em contextos de saúde (KUNDU, 2023). A escolha foi baseada no trabalho de Velpula e Sharma (2023). A ligação entre a ResNet e as camadas densas que farão o ajuste fino da classificação de glaucoma é feita através de uma camada de *GlobalAveragePooling*, que reduz a dimensionalidade do problema e também o número de parâmetros, sendo útil para a eficiência e generalização do modelo (KUMAR et al., 2021).

3.2.1.2 VGG-19

A VGG-19 é uma rede neural convolucional profunda desenvolvida pelo Visual Geometry Group, da Universidade de Oxford, e foi durante muito tempo um dos destaques em competições de

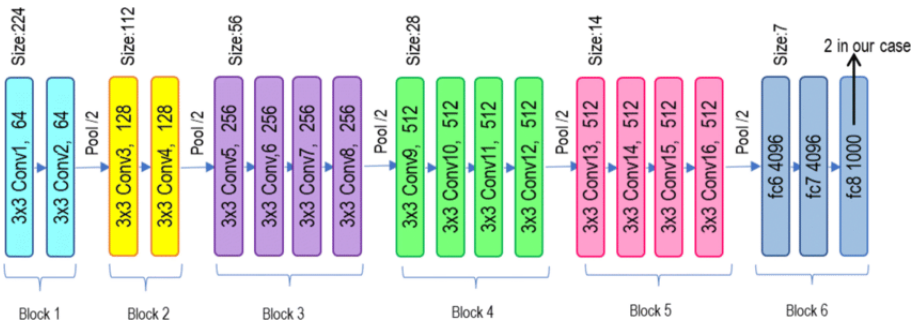
Figura 3.3 – Exemplo de uma conexão residual em uma ResNet.



Fonte: adaptado de (KUNDU, 2023)

reconhecimento e classificação de imagens, vencendo o ImageNet Challenge de 2014 (PANDEY, 2020). A principal melhoria aplicada por essa arquitetura foi aumentar significativamente a profundidade da rede, utilizando filtros de convolução pequenos (3×3), levando a um aumento substancial de performance em relação a modelos anteriores. A VGG-19, especificamente, contém 19 camadas de peso, sendo 16 camadas convolucionais e 3 camadas totalmente conectadas (SIMONYAN; ZISSERMAN, 2014) como pode ser observado na sua arquitetura representada na Figura 3.4. A escolha dessa arquitetura é baseada na sua recorrência em trabalhos de classificação de glaucoma (VELPULA; SHARMA, 2023), (ZHEN et al., 2018).

Figura 3.4 – Arquitetura básica da VGG



Fonte: (KHATTAR; QUADRI, 2022).

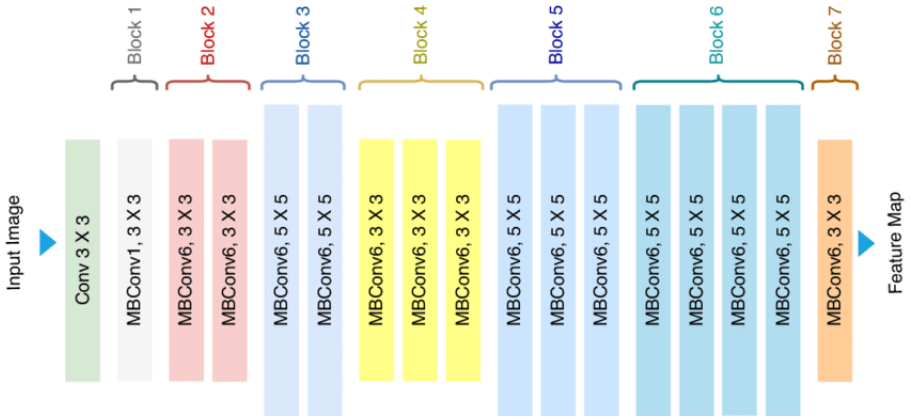
3.2.1.3 EffiencietNet-B0

A EfficientNet-B0 é uma rede neural convolucional de tamanho móvel que serve como a arquitetura base para a família de modelos EfficientNet. Diferente das redes anteriores, como ResNet e VGG, ela foi projetada para obter bom desempenho com menor número de parâmetros, tornando-a desejável para um teste que visa avaliar o balanço entre eficiência e desempenho (TAN; LE, 2019). Sua inovação está em um escalonamento composto, promovendo um aumento proporcional e equilibrado das dimensões de largura, profundidade e resolução, utilizando um coeficiente composto para garantir uniformidade.

A arquitetura da EfficientNet-B0 tem como base a MobileNetV2, incorporando elementos como o *Mobile Inverted Bottleneck Convolution* (SANDLER et al., 2018), uma unidade eficiente para dispositivos móveis. Outro fator incorporado e que contribui para a eficiência da rede é a otimização *squeeze-and-excitation*, um mecanismo que permite à rede aprender a importância de cada canal de características (HU; SHEN; SUN, 2018).

Na Figura 3.5 é possível observar a arquitetura básica da rede, com suas camadas mais detalhadas. Considerando sua capacidade de obter excelentes níveis de precisão com grande eficiência computacional (TAN; LE, 2019), torna-se viável explorar sua aplicação nos modelos deste trabalho, mesmo que não haja menções prévias da mesma na literatura para a classificação de glaucoma, como forma de se avaliar se uma solução mais barata computacionalmente pode ser viável.

Figura 3.5 – Arquitetura básica da EfficientNet-B0



Fonte: (AHMED; SABAB, 2022).

3.2.1.4 Demais camadas

Logo após, duas camadas densas totalmente conectadas foram utilizadas, visando refinar as características extraídas pelas convoluções realizadas nos modelos fundadores, cada uma com 256 e 128 neurônios, respectivamente, com a função de ativação ReLU. Essas camadas são intercaladas com *Batch Normalization*, que são inseridas com o objetivo de acelerar a convergência do modelo e, consequentemente, o treinamento como um todo. Isso ocorre devido à característica dessa técnica de reduzir a covariância interna (mudanças bruscas na distribuição dos dados ao longo das camadas). Dessa forma, é possível utilizar taxas de aprendizado maiores, como a utilizada no presente estudo (IOFFE; SZEGEDY, 2015).

Visando mitigar problemas com *overfitting*, foi aplicado *Dropout* logo após cada uma das camadas densas, desativando 30% dos neurônios a cada iteração entre cada uma das camadas. Aliado a ela, também há a aplicação de regularização L2, dessa forma, evitando que os pesos cresçam demais, e, consequentemente, fazendo com que os pesos menores distribuam melhor a importância das conexões (MOORE; DENERO, 2011).

Por fim, a camada final é composta por um neurônio, que irá ter como saída a probabilidade final. A função de ativação escolhida foi a *Sigmoid* devido a essa característica binária do problema. A função é útil pois apresenta o resultado final como uma única probabilidade contida entre 0 e 1, interpretada como a probabilidade de uma determinada instância ser da classe 1 ou não. Dessa forma, traçado um limiar de 0.5, é possível separar as classes de maneira fácil (KYURKCHIEV; MARKOV, 2015). A definição matemática da função foi observada na Equação 2.12.

3.2.2 *Vision Transformer*

Conforme apresentado na Seção 2.1.5.1, a arquitetura do ViT opera diretamente sobre partes da imagem e foi escolhido para utilização o modelo pré-treinado *google/vit-base-patch16-224-in21k*, disponibilizado pela *Hugging Face* [2]. Devido à essa disponibilização, várias funções e classes da biblioteca *transformers* foram utilizadas de maneira a facilitar a implementação, como a classe *Trainer*, que compacta diversos aspectos do treinamento, tornando o código mais direto e simples.

O modelo base foi pré-treinado de forma supervisionada no conjunto ImageNet-21k (cerca de 14 milhões de imagens e 21.843 classes), seguido de *fine-tuning* no ImageNet-1k (1 milhão de imagens divididas em 1.000 classes), ambos em resolução 224×224 *pixels*. As imagens são normalizadas com média $(0{,}5,\ 0{,}5,\ 0{,}5)$ e desvio padrão $(0{,}5,\ 0{,}5,\ 0{,}5)$, o que define a faixa de valores esperada pelo modelo.

Para adaptar esse modelo à tarefa de classificação entre Glaucoma/Normal, foi criada uma classe *ViTForImageClassification*, a qual permitiu substituir a camada de classificação original por uma camada linear ajustada ao número de classes do problema. Devido a essa camada possuir dois neurônios, uma função Softmax foi necessária, com sua fórmula apresentada na Equação (2.13), dando a probabilidade de cada uma das classes do problema. A estrutura interna do *encoder* foi mantida inalterada, garantindo o aproveitamento total dos pesos pré-treinados.

A seguir, uma síntese da metodologia do treinamento do ViT:

- **Modelo-base:** *google/vit-base-patch16-224-in21k*, com cabeça de classificação substituída por uma camada linear binária.
- **Pré-processamento:** transformações automáticas gerenciadas pelo `ViTImageProcessor`, incluindo redimensionamento para 224×224 e normalização conforme o esquema de pré-treinamento.
- **Principais hiperparâmetros de ajuste:**
  - ***Learning rate***: 1×10^{-5}, sendo um valor baixo pois não apresentou comportamento estável com valores maiores.

[2] Disponível em: <https://huggingface.co/google/vit-base-patch16-224>. Acesso em Junho de 2025.

  - ***Scheduler de learning rate***: `cosine_with_restarts`, com reinícios periódicos para evitar mínimos locais.
  - ***Batch size***: 16, compatível com a GPU disponível.
  - **Épocas**: 25 épocas
  - **Precisão mista**: ativação do modo `fp16` para acelerar o processo em GPU sem perder precisão numérica.

- **Métricas monitoradas**: acurácia, precisão, revocação, F1-score e AUC, todas calculadas a cada avaliação no conjunto de validação e posteriormente no conjunto de teste.

A escolha do `cosine_with_restarts` se dá devido a seu comportamento de variar a taxa de aprendizado seguindo uma curva cosseno, começando de um valor máximo e diminuindo gradualmente até um valor mínimo. Após atingir esse mínimo, a taxa de aprendizado reinicia e volta ao valor máximo, repetindo o processo. Isso previne quedas abruptas de desempenho e aumenta a chance de alcançar melhores mínimos (LOSHCHILOV; HUTTER, 2017).

3.2.3 MaxVit

Para complementar a análise com arquiteturas baseadas em *transformers*, também foi utilizado o modelo MaxViT, conforme descrito na Seção 2.1.5.2, através da implementação `maxvit_base_tf_384.in1k` [3] do repositório *timm*.

Diferente da abordagem com ViT, o MaxViT foi treinado manualmente em *PyTorch*, sem o uso de interfaces de alto nível como o `Trainer`, que é uma classe que automatiza todo o processo de treinamento, avaliação e salvamento de modelos, abstraindo os detalhes de treino e otimização. O modelo base foi carregado com pesos pré-treinados no conjunto ImageNet-1K e teve sua camada de classificação substituída por uma nova camada linear com duas unidades de saída, correspondentes às classes do problema, consequentemente, também utilizando uma função Softmax, descrita na Equação (2.13). Os pesos pré-treinados originais foram mantidos, assim como no ViT.

O começo da rede, devido à utilização de um modelo já pré-treinado, pede todos os dados em um formato específico, tornando todos os pré-processamentos descritos na Seção 3.1.1 necessários para o funcionamento e como primeiro passo do fluxo geral.

3.2.4 Esquema de votação

O algoritmo da votação proposto em si, começa com o carregamento e pré-processamento (redimensionamentos e normalizações) das imagens de forma similar ao descrito na Seção 3.1.1, ajustando-se assim à entrada de cada rede.

[3] Disponível em: <https://huggingface.co/timm/maxvit_base_tf_384.in1k>. Acesso em Julho de 2025.

Cada modelo então realiza suas predições individualmente, e essas probabilidades são agregadas para formar as saídas finais, como visto no Algoritmo 3.1. Com as probabilidades em mãos, ambas as votações são realizadas.

Algoritmo 3.1: Votação entre modelos

Input: Lista de caminhos das imagens X, Modelos M_i, Labels verdadeiros Y

Output: Predições finais por ambas votações

Carregar modelos CNNs, ViT e MaxViT;
Preprocessar as imagens conforme o modelo;
for cada modelo M_i **do**
    Realizar predição P_i com M_i sobre X;
Concatenar todas as predições em matriz P;
`// Votação por médias`
Calcular média das probabilidades $\bar{P} = mean(P, axis = 0)$;
Gerar predições $\hat{Y}_{mdias} = (\bar{P} > 0{,}5)$;
`// Votação por maiorias`
Converter P em binário: $P_b = (P > 0{,}5)$;
Gerar $\hat{Y}_{maioria} = round(mean(P_b, axis = 0))$;
Avaliar $\hat{Y}_{mdias}$ e $\hat{Y}_{maioria}$ usando Y;

Essa abordagem permite integrar diferentes perspectivas de extração de atributos (textura, padrões espaciais, relações globais) em uma única decisão, tendo apresentado resultados em (VELPULA; SHARMA, 2023), porém apenas com redes convolucionais e em bases reduzidas.

3.3 Avaliação

Para avaliação do desempenho, é importante olhar para métricas distintas, devido às suas diferentes aplicabilidades e interpretações. Como já demonstrado na Seção 2.1.10, a clássica acurácia não será uma métrica suficiente sozinha, dado o desequilíbrio entre classes presente no problema, embora seja importante observá-la principalmente para fins comparativos com outros trabalhos, além de ainda conseguir dar um panorama geral. De maneira geral, em cenários de saúde, a classe positiva para a enfermidade é muito menor que a classe negativa. Por exemplo, supondo que apenas 1% dos pacientes em um conjunto de dados tenha a doença, um modelo que sempre prevê "saudável" teria uma acurácia de 99%, mas seria completamente inútil, pois nunca detectaria a doença (LAVAZZA; MORASCA, 2023).

Posto isto, uma das métricas mais importantes a ser observada é o *recall*, visto que mede a capacidade do modelo de identificar corretamente todos os casos positivos, ou seja, pessoas que realmente têm a doença. Dessa forma, ao almejar uma boa métrica de *recall*, estamos indiretamente buscando minimizar os casos de falsos negativos, representando casos onde o modelo deixaria de dar um diagnóstico para um paciente doente (FILHO, 2023).

Embora o ponto de maior atenção seja essa identificação de falsos negativos, é desejável também que o modelo consiga uma precisão alta, isto é, tenha alta proporção de casos positivos identificados corretamente em relação a todos os casos classificados como positivos (HOSSIN; SULAIMAN, 2015). Isso pois, um *recall* muito alto com uma precisão baixa pode levar a muitos falsos positivos, o que pode sobrecarregar o sistema de saúde com exames desnecessários, por exemplo. Daí se faz importante um equilíbrio entre ambas as métricas, e para essa observação o *F1-Score* é útil por ser uma média harmônica de ambas, atingindo seu valor máximo (1) quando tanto a precisão quanto o *recall* são perfeitos, e é 0 quando o modelo falha completamente em uma das métricas (FILHO, 2023).

Por fim, o AUC também será observado, não apenas devido a fins comparativos, como também à sua interpretabilidade em contextos desbalanceados, medindo a capacidade de separação das classes independentemente de sua distribuição (RAKOTOMAMONJY, 2004).

4 Experimentos e Resultados

Neste capítulo, são detalhados tanto a configuração experimental aplicada, quanto os resultados obtidos. Na Seção 4.1 são apresentados detalhes pertinentes à realização e reprodução dos experimentos, tais como ambiente utilizado, versões de bibliotecas e escolhas de hiperparâmetros. Já os resultados obtidos, com uma discussão inicial de cada um deles, estão presentes na Seção 4.2, apresentando hipóteses e comparações visuais. Por fim, na Seção 4.3, uma discussão já mais completa e generalizada é apresentada, contendo demonstrações através de mapas de ativação.

4.1 Setup de experimentos

Para a realização dos experimentos, foi utilizada uma máquina com 16GB de memória RAM, processador Intel I7-13650HX e placa gráfica NVIDIA GEFORCE RTX 4060, em um ambiente Windows 11. Todos os códigos utilizados foram desenvolvidos utilizando o Python 3.10.11, e as principais bibliotecas utilizadas podem ser divididas em dois conjuntos, sendo um para as CNNs e outro para os modelos *transformers*.

- **Treinamento CNN**: As principais foram a *Tensorflow* e *Keras*, ambas na versão 2.10, além de outras utilitárias como Pandas (versão 2.2.3), NumPy (versão 1.26), Matplotlib (versão 3.9.2), Scikit-Learn (versão 1.5.2) e OpenCV (versão 4.10)
- **Treinamento *transformers***: principalmente o *torch* na versão 2.5.1+cu118, *timm* na 1.0.17, *transformers* na 4.52.3 e *huggingface-hub* na 0.32.1. Como auxiliares se têm NumPy (versão 2.1.2), Pandas (versão 2.2.3), Scikit-Learn (versão 1.6.1) e OpenCV (versão 4.11)
- **Ambiente para votação**: Como o ambiente de votação necessita tanto do *torch* quanto do *tensorflow* ao mesmo tempo, foi necessário um ajuste mais fino de versões para evitar problemas de compatibilidade. O *tensorflow* ficou na versão 2.13, enquanto o *torch* rodou na 2.1.2+cu118. As demais versões específicas poderão ser encontradas em um arquivo de requerimentos junto ao código fonte [1].

Seguem-se as escolhas pertinentes ao processo de treinamento do modelo, tais como otimizadores e taxas de aprendizado. Começando pelos modelos CNNs, para otimização, foi utilizado o RMSProp, começando com uma taxa de aprendizado de 1×10^{-4}. Ele é um otimizador adaptativo, ajustando a taxa de aprendizado de cada parâmetro individualmente com base na média móvel dos gradientes ao longo do tempo. Dessa forma, ele acaba se mostrando mais

[1] Disponível em: <https://github.com/ArthurCelestino/fundus_image_classification.git>.

estável, especialmente em redes profundas e problemas com gradientes ruidosos (ZOU et al., 2019).

Para a função de custo foi escolhida a já citada *Binary Crossentropy*, adequada para problemas como o apresentado neste trabalho, onde se tem apenas a classificação em doente ou não (RUBY; YENDAPALLI et al., 2020). Sua fórmula matemática foi apresentada previamente na Equação 2.4. Dessa forma, ela avalia quão bem o modelo está prevendo probabilidades para as classes 0 e 1. Se a previsão estiver correta e próxima de 1 para a classe positiva, ou de 0 para a classe negativa, a perda será pequena. Se a previsão estiver errada, a perda será maior, forçando o modelo a ajustar seus pesos durante o treinamento.

Agora, para as escolhas nos modelos ViT e MaxViT, na otimização, utilizou-se o AdamW com *learning rate* de 1×10^{-5} e 1×10^{-4}, respectivamente. Como visto na Seção 2.1.3.4, o AdamW é uma versão aprimorada do algoritmo Adam que é ideal para arquiteturas profundas e complexas como as baseadas em *transformers* (LOSHCHILOV; HUTTER, 2019), justificando assim sua escolha. Ele até mesmo é o otimizador padrão utilizado pela classe *Trainer*, da biblioteca *transformers*, utilizada no treinamento do ViT. O *Learning rate* maior do MaxViT deu-se aos *batches* de tamanho reduzido utilizados no treinamento do mesmo, visando evitar um ajuste muito justo a cada dado e, possivelmente, um *overfitting*.

Por fim, a função de custo nesses modelos *transformers* foi a também já citada *Categorical CrossEntropy*, devido à camada final destes modelos possuir dois neurônios, classificando-os como multiclasse. Essa função, como visto na Equação (2.5), mede o quão distante está a distribuição de probabilidade prevista da distribuição verdadeira e penaliza fortemente previsões confiantes, porém incorretas (PYKES, 2024).

4.2 Resultados

Nesta seção, são detalhados os resultados obtidos nos experimentos das arquiteturas individualmente e no resultado pós-aplicação da votação, com um foco comparativo visando descobrir qual estratégia se comporta melhor. O experimento foi realizado de forma que todos os modelos, juntamente com o algoritmo de votação, recebessem os mesmos dados de teste e treinamento, mantendo uma divisão com semente fixa desde o treinamento de cada um deles. Essa precaução foi necessária para evitar que dados usados para treinar qualquer um dos modelos pudessem estar presentes nos testes, assim os enviesando. No artigo que utilizou o esquema de votação originalmente (VELPULA; SHARMA, 2023), foi utilizado um *k-fold* com k = 5, porém, como se tratava de uma base bem menor e modelos mais simples, por questões práticas computacionais, os experimentos no presente trabalho foram realizados uma única vez. Nos demais trabalhos da literatura, não há muitas menções à quantidade de execuções realizadas. Por questões de tempo de execução e recursos computacionais, cada modelo foi treinado com quantidades de épocas e *batch size* diferentes, porém sempre deixando-os em treinamento até

se obter um valor pelo menos menor do que 0,1 de erro no conjunto de treinamento. No fim, as configurações destes dois parâmetros para cada modelo foram:

- **ResNet-50, VGG19 e EfficientNetB0**: 50 épocas com 16 de *batch size*
- **ViT**: 25 épocas com 32 de *batch size*
- **MaxVit**: 10 épocas com 4 de *batch size*

4.2.1 Resultados individuais CNNs

Primeiramente, para os modelos CNNs, avaliando o desempenho nos dados de teste, é possível ver na Tabela 4.1, que a EfficientNet-B0 apresentou resultados inferiores às demais, indicando que é demasiadamente simples para a tarefa e não conseguindo boa generalização. Uma das possíveis causas para tamanha diferença, além da simples diferença de robustez do modelo, pode ser o tamanho de entrada, onde está sendo utilizado 124×124, enquanto a rede utiliza normalmente 224×224, já que a família das arquiteturas EfficientNet foi construída justamente se baseando em mudanças de dimensões de redes, incluindo nisso a resolução de entrada, tendo diversas versões para justamente entradas distintas (TAN; LE, 2019). Dessa forma, ela será desconsiderada na posterior votação, e as análises se concentrarão nas duas outras redes. Entre as restantes, é observável que a ResNet50 apresentou resultados ligeiramente superiores em relação a VGG19 nas métricas avaliadas, com destaque para o valor de AUC (0,9339 contra 0,9257) e para a precisão (0,8904 contra 0,8721), indicando maior capacidade de discriminar corretamente as classes positivas e menor proporção de falsos positivos. O F1-score (0,8196 contra 0,8028) também sugere melhor equilíbrio entre precisão e *recall*.

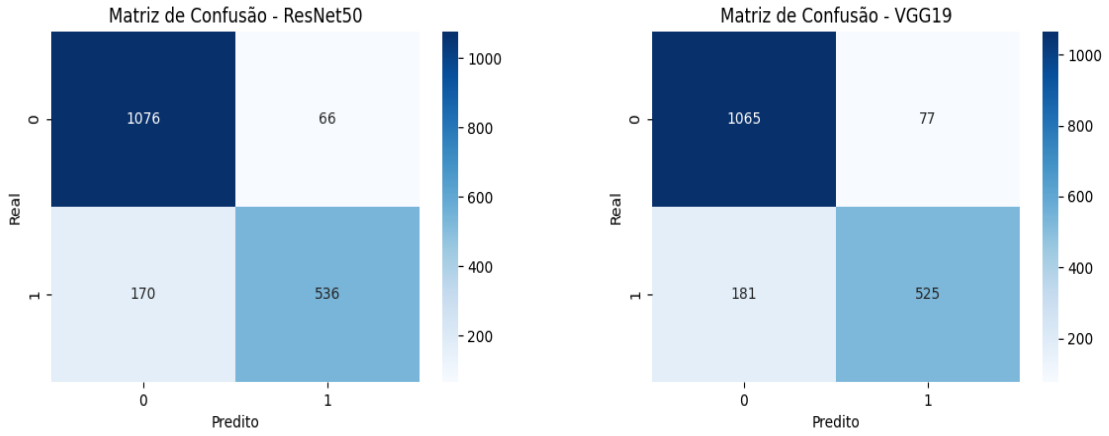
Tabela 4.1 – Resumo das métricas obtidas para ResNet50, VGG19 e EfficientNet-B0

Modelo	Acurácia	Precisão	Revocação	F1-score	AUC
ResNet50	0.8723	0.8904	0.7592	0.8196	0.9339
VGG19	0.8604	0.8721	0.7436	0.8028	0.9257
EfficientNet-B0	0.6996	0.5769	0.8016	0.6710	0.8218

Fonte: próprio autor.

Porém, de forma comum, estes dois modelos baseados em redes convolucionais clássicas apresentaram *recall* relativamente menor em relação à precisão, o que indica que ainda há margem para melhorias na detecção de casos positivos. Isso pode ser verificado na Figura 4.1, onde ambos apresentaram de maneira comum um valor maior de exemplos preditos na classe negativa, sendo que eram, na verdade, pertencentes à classe positiva, ou seja, FNs. Tal comportamento indica um dos piores cenários no contexto de detecção de doenças, onde o modelo deixaria de diagnosticar uma pessoa que possui a doença. Com isso, pode-se concluir que as arquiteturas convolucionais têm um desempenho mediano, acertando grande parte do que diz possuir o glaucoma, porém deixando passar casos, indicando que talvez não tenham conseguido captar e aprender elementos discriminativos nas imagens da forma desejada.

Figura 4.1 – Matrizes de confusão para cada um dos modelos, sendo (a) correspondente à ResNet-50 e (b) correspondente à VGG19. Relembrando que a classe 0 representa a classe Não-Glaucoma, enquanto a 1 represente a Glaucoma



(a) (b)

Fonte: próprio autor.

4.2.2 Resultados individuais *transformers*

Já para os resultados de ambos modelos baseados em *transformers*, as métricas, como observadas na Tabela 4.2, apresentaram comportamento diverso entre os dois modelos, indicando diferenças primordiais entre eles.

Tabela 4.2 – Resumo das métricas obtidas para ViT e MaxVit

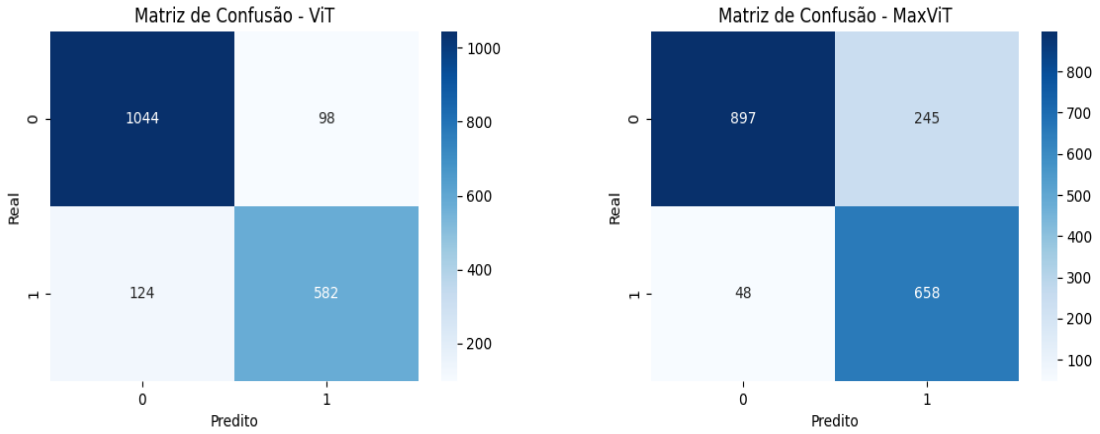
Modelos	Acurácia	Precisão	Revocação	F1-score	AUC
ViT	0.8799	0.8559	0.8244	0.8398	0.9294
MaxVit	0.8415	0.7287	0.9320	0.8179	0.9528

Fonte: próprio autor.

Observa-se que o ViT apresentou maior acurácia (0,8799 contra 0,8415) e precisão (0,8559 contra 0,7287), indicando que, proporcionalmente, ele comete menos falsos positivos, o que pode ser vantajoso para evitar alarmes indevidos em pacientes sem a doença, consequentemente também sobrecarregando o sistema de saúde. Por outro lado, o MaxViT obteve *recall* significativamente superior (0,9320 contra 0,8244), sugerindo maior capacidade de identificar casos positivos, deixando assim passar poucos casos de doentes erroneamente, ainda que com o custo de gerar mais falsos positivos, refletido na sua precisão mais baixa. Esse comportamento pode estar relacionado a diferenças estruturais entre as arquiteturas: enquanto o ViT segue um processamento puro via blocos de atenção, preservando maior seletividade na classificação, o MaxViT combina atenção local e global em diferentes escalas, o que pode ampliar a sensibilidade a padrões sutis nas imagens de fundo de olho, mas também aumentar a propensão a classificar amostras ambíguas como positivas. Apesar da menor acurácia geral, o MaxViT alcançou a maior AUC (0,9528 contra 0,9294), indicando melhor capacidade discriminativa global entre as classes.

Na prática, a escolha entre os modelos dependeria do contexto, sendo que no caso clínico, onde o principal objetivo seja única e exclusivamente minimizar o risco de não detectar um paciente com a doença, o MaxViT se mostra mais indicado. Porém, em países com sistemas de saúde cheios, como o Brasil (MASSUDA et al., 2018), o ViT apresenta um desempenho mais equilibrado, não gerando tantos alarmes falsos e podendo ser uma alternativa mais viável. Tal comportamento pode ser observado na Figura 4.2, onde tem-se uma clara distribuição quase que oposta nos setores da matriz (fora a diagonal principal onde se encontram os acertos).

Figura 4.2 – Matrizes de confusão para cada um dos modelos, sendo (a) correspondente ao ViT e (b) correspondente ao MaxViT.



(a) (b)

Fonte: próprio autor.

4.2.3 Resultados Votação

Finalizando os resultados, espera-se que o algoritmo de votação proposto consiga combinar valências dos diferentes modelos usados na execução do mesmo, para obter assim um desempenho equilibrado, porém ainda superior a qualquer modelo individualmente. Na Seção 3.2.4, se espera que as vantagens e desvantagens apresentadas no campo teórico se reflitam também nos testes.

Tabela 4.3 – Métricas obtidas tanto com a votação por média (*Soft Voting*), quanto com a por maioria (*Hard Voting*)

Esquema de votação	Acurácia	Precisão	Revocação	F1-score	AUC
Soft Voting	0.8945	0.8774	0.8414	0.8590	0.9567
Hard Voting	0.8874	0.9082	0.7847	0.8419	0.9567

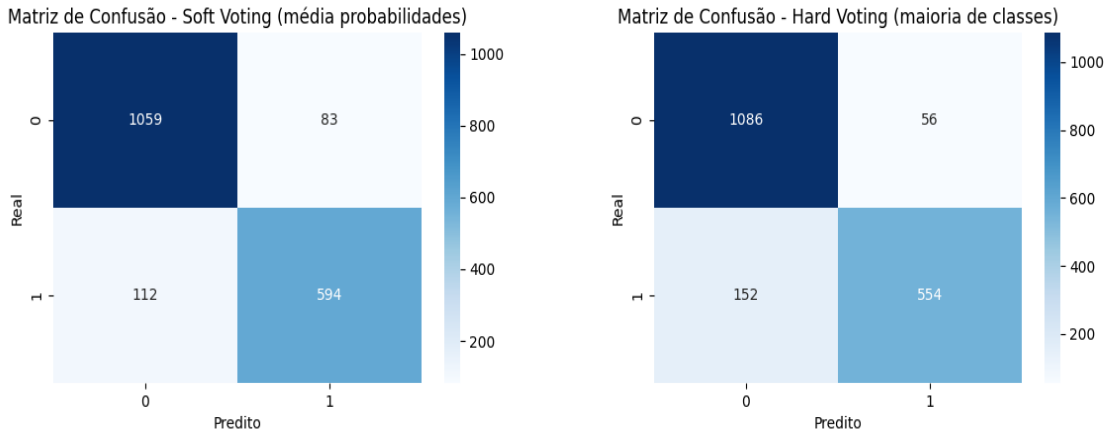
Fonte: próprio autor.

Como pode ser observado na Tabela 4.3, de maneira geral, a votação que levou em conta as médias de probabilidades obteve um resultado mais balanceado, sendo superior em quase todas as métricas, com exceção da precisão (0,8774 contra 0,9082), embora ainda esteja muito próxima. Como esperado, esse tipo de votação teve esse equilíbrio e superioridade, provavelmente

devido ao seu uso das informações probabilísticas dos modelos, especialmente dos modelos ViT e MaxViT, visto que individualmente demonstraram boas capacidades de generalização, puxando assim a média à favor dos mesmos. Esse bom balanço também pode ser explicado por um certo equilíbrio entre os modelos propostos, sem grandes *outliers* (mais de 20% de diferença entre alguma métrica específica), já que essa estratégia tende a ser mais sensível a modelos descalibrados. Com isso, o ponto forte da votação por maioria, discutido na Seção 3.2.4, acabou não sendo muito utilizado, deixando assim para ela apenas a desvantagem de não levar as probabilidades em conta.

Porém, nenhum dos dois esquemas alcançou um *recall* acima dos 0,85, demonstrando que alguns casos da doença ainda são perdidos, como visto nas matrizes de confusão expostas na Figura 4.3. Tal comportamento pode ser explicado pelo fato de que apenas o MaxViT conseguiu um *recall* acima dos 0,90 dentre todos os modelos, destoando-se talvez um pouco dos demais, tornando-o assim voto vencido na votação por maioria e também não sendo suficiente sozinho para subir a média no outro esquema nas imagens doentes perdidas pelo algoritmo.

Figura 4.3 – Matrizes de confusão para cada esquema, com (a) sendo do *Soft Voting* e (b) do *Hard Voting*.



(a) (b)

Fonte: próprio autor.

4.3 Discussão dos Resultados

Nesta seção, a partir dos resultados apresentados individualmente na seção anterior, os valores foram reunidos na Tabela 4.4, de forma a possibilitar uma comparação direta entre todos os métodos propostos neste trabalho.

Observa-se que os modelos puramente convolucionais não se destacaram em nenhuma métrica específica, confirmando que, no geral, ficam atrás das arquiteturas mais complexas baseadas em *transformers*. O ponto mais crítico desses modelos está na baixa capacidade de identificar casos de glaucoma, refletida no reduzido *recall* das CNNs, ainda que sejam amplamente

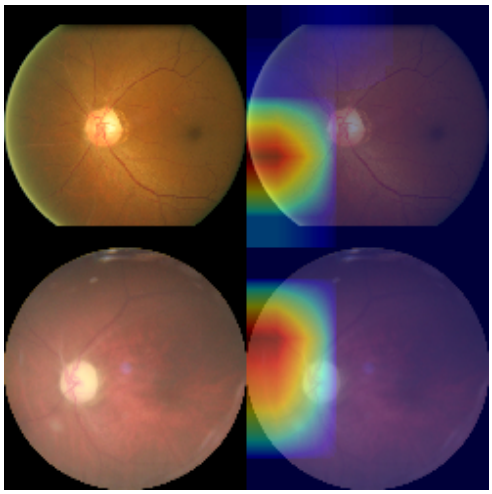
Tabela 4.4 – Resumo comparativo de métricas por modelo/algoritmo

Modelo/Algoritmo	**Acurácia**	**Precisão**	***Recall***	**F1-score**	**AUC**
ResNet50	0.8723	0.8904	0.7592	0.8196	0.9339
VGG19	0.8604	0.8721	0.7436	0.8028	0.9257
EffNetB0	0.6996	0.5769	0.8016	0.6710	0.8218
ViT	0.8799	0.8559	0.8244	0.8398	0.9294
MaxVit	0.8415	0.7287	**0.9320**	0.8179	0.9528
Soft Voting	**0.8945**	0.8774	0.8414	**0.8590**	**0.9567**
Hard Voting	0.8874	**0.9082**	0.7847	0.8419	**0.9567**

Fonte: próprio autor.

utilizadas na literatura, sobretudo antes da popularização de arquiteturas como o ViT. Buscando entender visualmente tal aspecto da métrica, na Figura 4.4 é possível ver que modelos como a ResNet focaram diversas vezes em locais da imagem que não exatamente o OD, região onde estão concentradas as características discriminatórias da enfermidade.

Figura 4.4 – Mapa de ativação de dois FN da ResNet50.

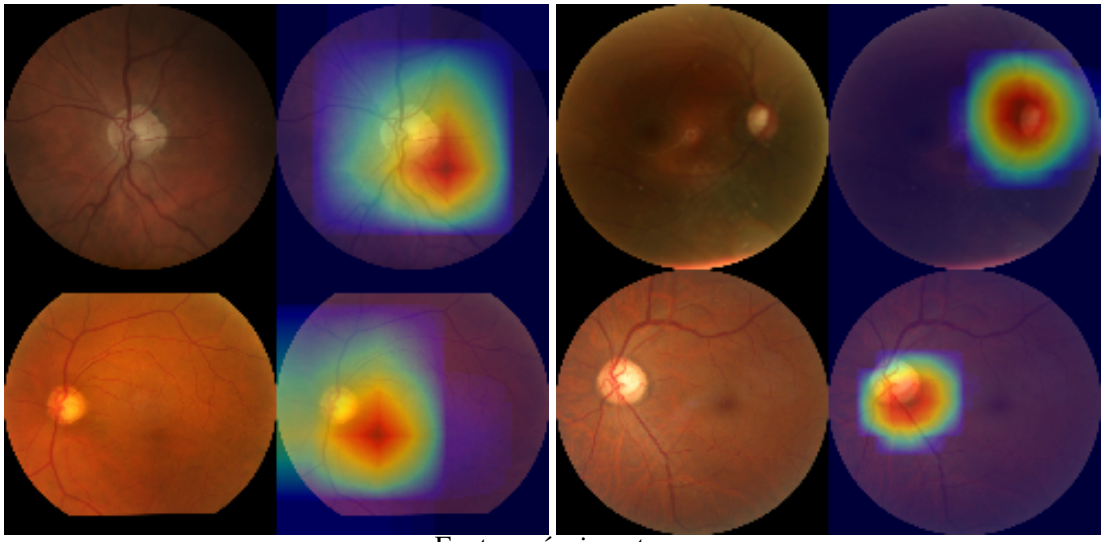


Fonte: próprio autor.

Já para casos de FPs, os mapas de ativação até mostram foco nas regiões corretas, como visto na Figura 4.5 , com a ResNet levemente mais deslocada do que a VGG, porém provavelmente acabaram se prendendo a aspectos dessas regiões que não representavam a doença bem.

Por outro lado, soluções como o ViT base e o MaxViT apresentaram desempenhos superiores na revocação, considerada primordial no contexto clínico. Em especial, o MaxViT atingiu *recall* de 0,93, evidenciando excelente capacidade de detecção de amostras com a doença. No entanto, essa performance veio acompanhada de baixa precisão, o que indica um *trade-off* importante: ao aumentar a sensibilidade para capturar mais casos positivos, ele também eleva o número de falsos positivos. Essa característica, embora aceitável ou até desejável em cenários onde não se pode correr o risco de deixar pacientes doentes passarem despercebidos, pode gerar sobrecarga em sistemas de saúde com recursos limitados, devido ao retrabalho exigido para confirmar diagnósticos. Os mapas de ativação, conforme mostrados na Figura 4.6, mostram

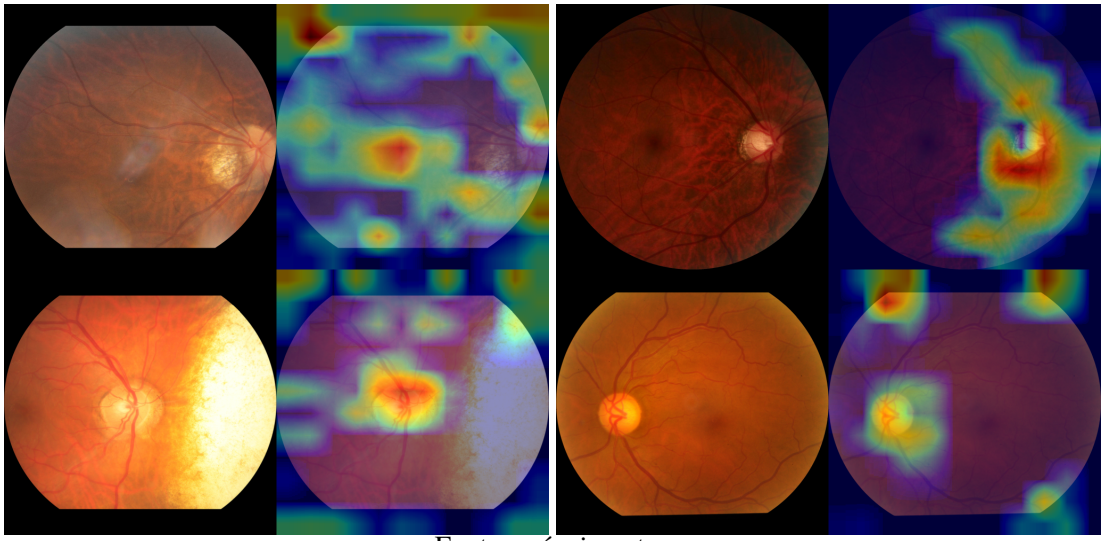
Figura 4.5 – Mapas de ativação de falsos positivos da ResNet50 à esquerda, e da VGG19 à direita.



Fonte: próprio autor.

que a característica de olhar um contexto mais global acabou prejudicando nesse excesso de diagnósticos, com mapas de FPs ativados em diversas regiões aleatórias da imagem. Já nos FNs, nota-se que mesmo errando, ele acabou se atentando às regiões minimamente corretas, sendo um bom indicativo da forma como o modelo foca sua atenção majoritária.

Figura 4.6 – Mapas de ativação do MaxViT, sendo os falsos positivos representados à esquerda e os falsos negativos à direita.

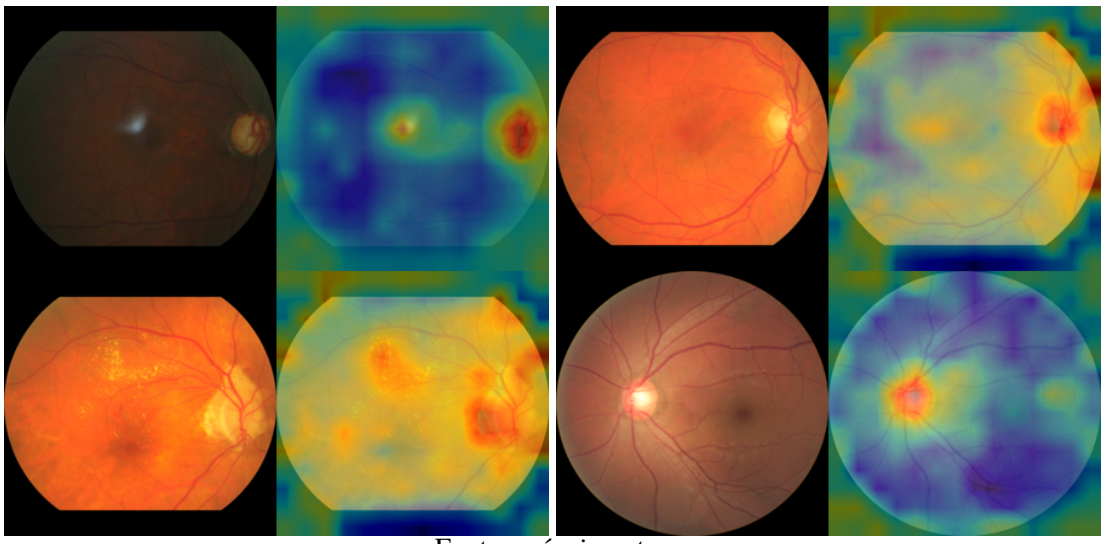


Fonte: próprio autor.

Já o ViT apresentou valores mais equilibrados entre precisão e recall, entregando um desempenho bom e com menor perda de um dos lados dessa balança. Porém, esse bom balanço se torna de difícil explicabilidade, dado que os *patches* não têm relações globais fortes entre si como no MaxViT. Na Figura 4.7 esse fator é percebido com ativações espalhadas por toda a imagem. Porém, mesmo com esse aspecto, é notável que a região do OD ainda costuma ter mais destaque, mesmo em FNs. Esse equilíbrio das demais métricas é refletido também no AUC, onde tanto o

ViT quanto o MaxViT obtiveram valores próximos (0,9294 e 0,9528, respectivamente), indicando boa separabilidade entre as classes. Ainda assim, o AUC mais alto do MaxViT confirma que, apesar da baixa precisão, ele mantém um poder de discriminação elevado.

Figura 4.7 – Mapas de ativação do ViT, com os falsos positivos representados na esquerda e os falsos negativos na direita



Fonte: próprio autor.

Ao analisar o conjunto de todos os modelos e o sistema de votação, nota-se que o *Soft Voting* apresentou o desempenho mais equilibrado entre todas as soluções. Contudo, seu uso implica maior custo computacional e tempo de processamento, já que cada inferência exige o carregamento e execução de quatro modelos. Considerando que, em muitas métricas, o ganho foi de aproximadamente 2% em relação ao ViT (segundo modelo mais equilibrado), é necessário ponderar a adoção dessa abordagem conforme o contexto de aplicação.

Por exemplo, em cenários onde evitar falsos negativos seja a prioridade, a utilização isolada do MaxViT pode ser vantajosa. Já em situações de maior demanda e necessidade de processamento mais ágil, o ViT pode ser uma escolha mais apropriada, reduzindo o tempo de inferência sem comprometer significativamente o desempenho.

Tendo em vista estes pontos da análise, nem todos os modelos serão comparados com a literatura, excluindo-se os convolucionais que não se sobressaíram em nenhuma métrica e dando-se mais atenção às estratégias de votação e aos modelos *transformers*, dados seus resultados superiores no presente trabalho. Na Tabela 4.5, estão os resultados comparativos dos modelos escolhidos contra trabalhos da literatura que utilizam o SMDG-19 como conjunto de dados.

Com base nos resultados da Tabela 4.5, é possível observar que os modelos propostos ainda necessitam de melhorias em termos de desempenho bruto, principalmente comparado ao trabalho de Tohye et al. (2024). Foi utilizada a mesma proporção na divisão de dados, porém, mesmo assim, há diferenças na casa de 8-10% em algumas métricas para a maioria dos modelos, especialmente no *recall*, uma das principais a ser observada. Embora o estudo de Tohye et al. (2024) tenha

Tabela 4.5 – Comparativo dos modelos baseados em *transformers* e sistemas de votação com trabalhos da literatura utilizando a mesma base de dados (KIEFER, 2023).

Modelo	**Acurácia (%)**	**Precisão (%)**	***Recall* (%)**	**F1-Score**	**AUC**
ViT	87,99	85,59	82,44	0,8398	0,9294
MaxViT	84,15	72,87	**93,20**	0,8179	0,9528
Soft Voting	**89,45**	87,74	84,14	**0,8590**	**0,9567**
Hard Voting	88,74	**90,82**	78,47	0,8419	**0,9567**
Dosovitskiy et al. (2021)	-	-	-	-	0,95
Tohye et al. (2024)	93,00	93,00	93,08	0,9290	-

Fonte: próprio autor.

alcançado acurácia e precisão superiores (93% em ambas), o *Soft Voting* desenvolvido apresentou diferenças de apenas 5%, com um AUC bem elevado de 0,9567, indicando boa capacidade de discriminação entre as classes e revelando um potencial de melhoria caso os modelos utilizados consigam captar melhor informações discriminativas. Além disso, o MaxViT destacou-se pelo *recall* de 93,20%, superando inclusive os trabalhos de referência nessa métrica, o que reforça seu potencial para aplicações onde a detecção de todos os casos positivos seja prioritária. Já em relação ao trabalho de Dosovitskiy et al. (2021), embora não tenham todas as métricas clássicas (como acurácia, precisão, *recall* e F1-score) de classificação disponíveis, é possível perceber que o AUC da grande parte dos modelos se mostrou equiparado, indicando uma boa separação entre as classes.

Conclui-se dessa forma que os modelos propostos ainda estão realmente inferiores à literatura, porém ainda possuem espaço para melhoria, dado que já apresentam valias interessantes em aspectos específicos, com um bom potencial de separação entre as classes.

5 Considerações Finais

Neste capítulo serão apresentadas conclusões gerais acerca do que já foi apresentado e dos experimentos finais na Seção 5.1, além de salientar o progresso em relação aos objetivos traçados no Capítulo 1.

5.1 Conclusão

O presente trabalho teve como objetivo principal a construção e avaliação de um conjunto de modelos baseados em redes neurais convolucionais e *transformers* para a classificação de glaucoma em imagens de fundo de olho, com foco na junção de suas previsões para mitigar erros individuais. Ao longo deste estudo, foi explorado o desempenho de diversas arquiteturas, tanto individualmente quanto combinadas, utilizando a base de dados SMDG-19, concluindo assim todas as etapas traçadas inicialmente. Os experimentos iniciais, abordando os modelos individualmente, forneceram *insights* importantes:

- As CNNs clássicas, como ResNet-50 (acurácia de 87,23%, *recall* de 75,92%) e VGG-19 (acurácia de 86,04%, recall de 74,36%), apresentaram resultados consistentes, mas com uma capacidade de detecção de casos positivos inferior aos *transformers*. A EfficientNet-B0, embora mais eficiente computacionalmente, mostrou-se demasiadamente simples para a complexidade da tarefa, com desempenho significativamente menor (acurácia de 69,96%).
- Entre os modelos *transformer*, o MaxViT destacou-se com um *recall* notável de 93,20% e o maior AUC de 0,9528. Essa característica é crucial em contextos clínicos, onde minimizar falsos negativos (não detectar um paciente doente) é a prioridade. Contudo, essa alta sensibilidade veio acompanhada de menor precisão. Por outro lado, o ViT apresentou um desempenho mais equilibrado (acurácia de 87,99%, precisão de 85,59% e recall de 82,44%), sendo uma opção viável em cenários que exigem um balanço entre precisão e sensibilidade.

A investigação central do trabalho, que consistia na utilização de esquemas de votação entre modelos diversos, demonstrou ser uma abordagem com suas vantagens e desvantagens, porém com resultados interessantes principalmente se for levado em conta a utilização de um *dataset* bem mais robusto do que os normalmente utilizados. O *Soft Voting* (votação por média), que considera as probabilidades de cada modelo, alcançou o melhor desempenho geral e mais equilibrado, com acurácia de 89,45%, precisão de 87,74%, *recall* de 84,14%, F1-score de 0,8590 e AUC de 0,9567. Esse resultado confirma que a combinação de diferentes arquiteturas, que capturam padrões de maneira distinta, pode superar as limitações de modelos individuais,

respondendo afirmativamente à pergunta de pesquisa proposta. O *Hard Voting*, por sua vez, obteve resultados ligeiramente inferiores ao concorrente em quase todas as métricas, exceto precisão, por desconsiderar o grau de confiança das previsões. Um ponto de atenção em comum para ambas as votações é seu aumento do tempo de processamento, com a utilização simultânea de vários modelos, levantando um ponto a ser analisado de acordo com o contexto prático de aplicação.

Apesar dos avanços alcançados, especialmente com o *Soft Voting* e o *recall* do MaxViT, a comparação com a literatura recente, como o trabalho de Tohye et al. (2024), utilizando a mesma base de dados, indica que ainda há espaço para melhorias significativas no desempenho bruto (com diferenças de acurácia de cerca de 5% e *recall* similar, exceto o MaxViT isolado).

Como trabalho futuro, sugere-se aprofundar a abordagem explorando diferentes técnicas de pré-processamento e segmentação, presentes em diversos trabalhos da literatura, sendo que esse ponto pode ajudar a isolar a região do OD, onde estão presentes os maiores sinais do glaucoma. Além disso, futuras pesquisas podem investigar métodos mais avançados de fusão de modelos, a otimização dos pesos de cada modelo no esquema de votação, ou a aplicação de técnicas de *explainable AI* para aumentar a interpretabilidade dos resultados em um contexto clínico. Com a continuidade dessas investigações enquanto sugestões de trabalhos futuros, espera-se contribuir para o desenvolvimento de soluções mais precisas e robustas para o diagnóstico precoce do glaucoma.

Referências

AGGARWAL, C. C. et al. Neural networks and deep learning. [S.l.]: Springer, 2018. v. 10.

AHMED, T.; SABAB, N. Classification and understanding of cloud structures via satellite images with efficientunet. SN Computer Science, v. 3, 01 2022.

ALLISON, K. et al. Primary open angle glaucoma: Where are we today? EC Ophthalmology, v. 15, p. 01–15, 2024.

ALMEIDA, M. Machine Learning: o que é aprendizado semi-supervisionado | Alura — alura.com.br. 2023. <https://www.alura.com.br/artigos/machine-learning-aprendizado-semi-supervisionado?srsltid=AfmBOorT_YMbG9zUFJm0beagbo8vOupyh6zDIbeAVcvTRJpxNxDK7ihD>. [Accessed 30-11-2024].

ALPAYDIN, E. Machine learning. [S.l.]: MIT press, 2021.

ARAúJO, J.; PAIVA, A. de; ALMEIDA, J. de; NETO, O. P. S.; SOUSA, J. de; SILVA, A.; JúNIOR, G. B. Diagnóstico de glaucoma em imagens de fundo de olho utilizando os Índices de diversidade de shannon e mcintosh. In: Anais do XVII Workshop de Informática Médica. Porto Alegre, RS, Brasil: SBC, 2017. p. 1873–1882. ISSN 2763-8952. Disponível em: <https://sol.sbc.org.br/index.php/sbcas/article/view/3698>.

ASHTARI-MAJLAN, M.; DEHSHIBI, M. M.; MASIP, D. Deep learning and computer vision for glaucoma detection: A review. arXiv preprint arXiv:2307.16528, 2023.

AWOFESO, Z. An Explanation of the Vision Transformer (ViT) Paper. Medium, 2024. Disponível em: <https://medium.com/codex/an-explanation-of-the-vision-transformer-vit-paper-8cdd399741aa>.

BANG, J.-H.; PARK, S.-W.; KIM, J.-Y.; PARK, J.; HUH, J.-H.; JUNG, S.-H.; SIM, C.-B. Ca-cmt: Coordinate attention for optimizing cmt networks. IEEE Access, IEEE, v. 11, p. 76691–76702, 2023.

BHATT, D.; PATEL, C.; TALSANIA, H.; PATEL, J.; VAGHELA, R.; PANDYA, S.; MODI, K.; GHAYVAT, H. Cnn variants for computer vision: History, architecture, application, challenges and future scope. Electronics, v. 10, n. 20, 2021. ISSN 2079-9292. Disponível em: <https://www.mdpi.com/2079-9292/10/20/2470>.

BJORCK, N.; GOMES, C. P.; SELMAN, B.; WEINBERGER, K. Q. Understanding batch normalization. Advances in neural information processing systems, v. 31, 2018.

BRAGANÇA, C. P.; TORRES, J. M.; SOARES, C. P. de A. Inteligência artificial e diagnóstico do glaucoma. Brazilian Applied Science Review, v. 7, n. 2, p. 683–707, 2023.

BRAUWERS, G.; FRASINCAR, F. A general survey on attention mechanisms in deep learning. IEEE Transactions on Knowledge and Data Engineering, Institute of Electrical and Electronics Engineers (IEEE), v. 35, n. 4, p. 3279–3298, abr. 2023. ISSN 2326-3865. Disponível em: <http://dx.doi.org/10.1109/TKDE.2021.3126456>.

BRUNETTE, E. S.; FLEMMER, R. C.; FLEMMER, C. L. A review of artificial intelligence. In: IEEE. 2009 4th International Conference on Autonomous Robots and Agents. [S.l.], 2009. p. 385–392.

CHEN, X.; XU, Y.; YAN, S.; WONG, D. W. K.; WONG, T. Y.; LIU, J. Automatic feature learning for glaucoma detection based on deep learning. In: SPRINGER. Medical Image Computing and Computer-Assisted Intervention–MICCAI 2015: 18th International Conference, Munich, Germany, October 5-9, 2015, Proceedings, Part III 18. [S.l.], 2015. p. 669–677.

CHINCHOLI, F.; KOESTLER, H. Transforming glaucoma diagnosis: transformers at the forefront. Frontiers in Artificial Intelligence, v. 7, 2024. ISSN 2624-8212. Disponível em: <https://www.frontiersin.org/journals/artificial-intelligence/articles/10.3389/frai.2024.1324109>.

CHOI, D.; SHALLUE, C. J.; NADO, Z.; LEE, J.; MADDISON, C. J.; DAHL, G. E. On empirical comparisons of optimizers for deep learning. arXiv preprint arXiv:1910.05446, 2019.

DOSOVITSKIY, A.; BEYER, L.; KOLESNIKOV, A.; WEISSENBORN, D.; ZHAI, X.; UNTERTHINER, T.; DEHGHANI, M.; MINDERER, M.; HEIGOLD, G.; GELLY, S. et al. An image is worth 16x16 words: Transformers for image recognition at scale. arXiv preprint arXiv:2010.11929, 2020.

DOSOVITSKIY, A.; BEYER, L.; KOLESNIKOV, A.; WEISSENBORN, D.; ZHAI, X.; UNTERTHINER, T.; DEHGHANI, M.; MINDERER, M.; HEIGOLD, G.; GELLY, S.; USZKOREIT, J.; HOULSBY, N. An Image is Worth 16x16 Words: Transformers for Image Recognition at Scale. 2021. Disponível em: <https://arxiv.org/abs/2010.11929>.

ELSHAMY, R.; ABU-ELNASR, O.; ELHOSENY, M.; ELMOUGY, S. Improving the efficiency of rmsprop optimizer by utilizing nestrove in deep learning. Scientific Reports, Nature Publishing Group UK London, v. 13, n. 1, p. 8814, 2023.

ERDOğAN, A. Squeeze-and-Excitation Networks. 2022. <https://medium.com/@atakanerdogan305/squeeze-and-excitation-networks-c4e1ad7d8a3d>. Publicado em 16 de outubro de 2022; acessado em 20 de agosto de 2025.

FACE, H. Vision Transformer (ViT) — transformers v4.11.1 documentation. 2021. Acessado em: 03 jul. 2025. Disponível em: <https://huggingface.co/transformers/v4.11.1/model_doc/vit.html>.

FERNEDA, E. Redes neurais e sua aplicação em sistemas de recuperação de informação. Ciência da Informação, SciELO Brasil, v. 35, p. 25–30, 2006.

FILHO, M. Precisão, recall e f1 score em machine learning. 2023. Disponível em: <https://towardsdatascience.com/a-look-at-precision-recall-and-f1-score-36b5fd0dd3ec>.

GAO, B.; PAVEL, L. On the Properties of the Softmax Function with Application in Game Theory and Reinforcement Learning. 2018. Disponível em: <https://arxiv.org/abs/1704.00805>.

GeeksforGeeks. Voting in Machine Learning. 2025. <https://www.geeksforgeeks.org/machine-learning/voting-in-machine-learning/>. Última atualização em 06 de agosto de 2025; acesso em 19 de agosto de 2025.

GHOLAMALINEZHAD, H.; KHOSRAVI, H. Pooling methods in deep neural networks, a review. arXiv preprint arXiv:2009.07485, 2020.

HOSSIN, M.; SULAIMAN, M. N. A review on evaluation metrics for data classification evaluations. International journal of data mining & knowledge management process, Academy & Industry Research Collaboration Center (AIRCC), v. 5, n. 2, p. 1, 2015.

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.

HWANG, E. E.; CHEN, D.; HAN, Y.; JIA, L.; SHAN, J. Multi-dataset comparison of vision transformers and convolutional neural networks for detecting glaucomatous optic neuropathy from fundus photographs. Bioengineering, v. 10, n. 11, 2023. ISSN 2306-5354. Disponível em: <https://www.mdpi.com/2306-5354/10/11/1266>.

IDREES, H. Vision Transformer vs CNN: A Comparison of Two Image Processing Giants. Medium, 2024. Disponível em: <https://medium.com/@hassaanidrees7/vision-transformer-vs-cnn-a-comparison-of-two-image-processing-giants-d6c85296f34f>.

IOFFE, S.; SZEGEDY, C. Batch Normalization: Accelerating Deep Network Training by Reducing Internal Covariate Shift. 2015. Disponível em: <https://arxiv.org/abs/1502.03167>.

ISLAM, S.; ELMEKKI, H.; ELSEBAI, A.; BENTAHAR, J.; DRAWEL, N.; RJOUB, G.; PEDRYCZ, W. A comprehensive survey on applications of transformers for deep learning tasks. Expert Systems with Applications, Elsevier, v. 241, p. 122666, 2024.

JANIESCH, C.; ZSCHECH, P.; HEINRICH, K. Machine learning and deep learning. Electronic Markets, Springer, v. 31, n. 3, p. 685–695, 2021.

KHATTAR, A.; QUADRI, S. "generalization of convolutional network to domain adaptation network for classification of disaster images on twitter". Multimedia Tools and Applications, v. 81, 09 2022.

KIEFER, R. SMDG, A Standardized Fundus Glaucoma Dataset. Kaggle, 2023. Disponível em: <https://www.kaggle.com/ds/2329670>.

KIEFER, R.; ABID, M.; STEEN, J.; ARDALI, M. R.; AMJADIAN, E. A catalog of public glaucoma datasets for machine learning applications: A detailed description and analysis of public glaucoma datasets available to machine learning engineers tackling glaucoma-related problems using retinal fundus images and oct images. In: Proceedings of the 2023 7th International Conference on Information System and Data Mining. New York, NY, USA: Association for Computing Machinery, 2023. (ICISDM '23), p. 24–31. ISBN 9798400700637. Disponível em: <https://doi.org/10.1145/3603765.3603779>.

KIM, H. E.; COSA-LINAN, A.; SANTHANAM, N.; JANNESARI, M.; MAROS, M. E.; GANSLANDT, T. Transfer learning for medical image classification: a literature review. BMC medical imaging, Springer, v. 22, n. 1, p. 69, 2022.

KOSSAIFI, J.; TOISOUL, A.; BULAT, A.; PANAGAKIS, Y.; HOSPEDALES, T. M.; PANTIC, M. Factorized higher-order cnns with an application to spatio-temporal emotion estimation. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. [S.l.: s.n.], 2020. p. 6060–6069.

KRIžAJ, P. D. What is glaucoma? 2019. <https://www.ncbi.nlm.nih.gov/books/NBK543075/>. [Accessed 30-11-2024].

KUMAR, D. Max pooling. Medium, 2023. Disponível em: <https://medium.com/@danushidk507/max-pooling-ef545993b6e4>.

KUMAR, R. L.; KAKARLA, J.; ISUNURI, B. V.; SINGH, M. Multi-class brain tumor classification using residual network and global average pooling. Multimedia Tools and Applications, Springer, v. 80, n. 9, p. 13429–13438, 2021.

KUNDU, N. Exploring Resnet50: An in-depth look at the model architecture and code implementation. Medium, 2023. Disponível em: <https://medium.com/@nitishkundu1993/exploring-resnet50-an-in-depth-look-at-the-model-architecture-and-code-implementation-d8d8fa67e46f>.

KYURKCHIEV, N.; MARKOV, S. Sigmoid functions: some approximation and modelling aspects. LAP LAMBERT Academic Publishing, Saarbrucken, v. 4, p. 34, 2015.

LAVAZZA, L.; MORASCA, S. Common problems with the usage of f-measure and accuracy metrics in medical research. IEEE Access, IEEE, v. 11, p. 51515–51526, 2023.

LECUN, Y.; BENGIO, Y.; HINTON, G. Deep learning. nature, Nature Publishing Group UK London, v. 521, n. 7553, p. 436–444, 2015.

LI, Z.; HE, Y.; KEEL, S.; MENG, W.; CHANG, R. T.; HE, M. Efficacy of a deep learning system for detecting glaucomatous optic neuropathy based on color fundus photographs. Ophthalmology, Elsevier, v. 125, n. 8, p. 1199–1206, 2018.

LOSHCHILOV, I.; HUTTER, F. SGDR: Stochastic Gradient Descent with Warm Restarts. 2017. Disponível em: <https://arxiv.org/abs/1608.03983>.

LOSHCHILOV, I.; HUTTER, F. Decoupled Weight Decay Regularization. 2019. Disponível em: <https://arxiv.org/abs/1711.05101>.

LUWEI, X.; HU, X.; CHEN, Y.; XUE, Y.; CHEN, B.; GU, D.; TANG, B. Multi-head self-attention based gated graph convolutional networks for aspect-based sentiment classification. Multimedia Tools and Applications, v. 81, p. 1–20, 06 2022.

MAHARANA, K.; MONDAL, S.; NEMADE, B. A review: Data pre-processing and data augmentation techniques. Global Transitions Proceedings, Elsevier, v. 3, n. 1, p. 91–99, 2022.

MARTIN, P.-E.; BENOIS-PINEAU, J.; PÉTERI, R.; MORLIER, J. Sport action recognition with siamese spatio-temporal cnns: Application to table tennis. In: IEEE. 2018 International conference on content-based multimedia indexing (CBMI). [S.l.], 2018. p. 1–6.

MASSUDA, A.; HONE, T.; LELES, F. A. G.; CASTRO, M. C. D.; ATUN, R. The brazilian health system at crossroads: progress, crisis and resilience. BMJ global health, BMJ Publishing Group Ltd, v. 3, n. 4, 2018.

MITCHELL, T. Machine Learning. McGraw-Hill, 1997. (McGraw-Hill International Editions). ISBN 9780071154673. Disponível em: <https://books.google.com.br/books?id=EoYBngEACAAJ>.

MOORE, R.; DENERO, J. L1 and l2 regularization for multiclass hinge loss models. In: Symposium on machine learning in speech and language processing. [S.l.: s.n.], 2011.

MORID, M. A.; BORJALI, A.; FIOL, G. D. A scoping review of transfer learning research on medical image analysis using imagenet. Computers in biology and medicine, Elsevier, v. 128, p. 104115, 2021.

MUMUNI, A.; MUMUNI, F. Data augmentation: A comprehensive survey of modern approaches. Array, Elsevier, v. 16, p. 100258, 2022.

MUREL, J.; KAVLAKOGLU, E. O que É aprendizado por transferência? 2025. Disponível em: <https://www.ibm.com/br-pt/think/topics/transfer-learning>.

OH, S.; PARK, Y.; CHO, K. J.; KIM, S. J. Explainable machine learning model for glaucoma diagnosis and its interpretation. Diagnostics, v. 11, n. 3, 2021. ISSN 2075-4418. Disponível em: <https://www.mdpi.com/2075-4418/11/3/510>.

PANDEY, R. P. Image classifier using VGG-19 deep learning model in google colab notebook. dishes detection. Medium, 2020. Disponível em: <https://medium.com/@ravipandey71998/image-classifier-using-vgg-19-deep-learning-model-in-google-colab-notebook-dishes-detection-34861168e0

POOJARY, R.; RAINA, R.; KRISHANMURTHY, S. Application of cnns in home security. In: IEEE. 2022 International Conference on Electrical and Computing Technologies and Applications (ICECTA). [S.l.], 2022. p. 322–327.

PYKES, K. Cross-Entropy Loss Function in Machine Learning: Enhancing Model Accuracy. 2024. Acesso em: 8 de agosto de 2025. Disponível em: <https://www.datacamp.com/tutorial/the-cross-entropy-loss-function-in-machine-learning>.

QUIGLEY, H. A. Open-angle glaucoma. New England Journal of Medicine, Mass Medical Soc, v. 328, n. 15, p. 1097–1106, 1993.

RAJ, U. Dropping the Knowledge Bomb: Understanding Dropout Layers in Deep Learning — utsavraj.ptn04. 2023. <https://medium.com/@utsavraj.ptn04/dropping-the-knowledge-bomb-understanding-dropout-layers-in-deep-learning-0612f517269d>. [Accessed 06-12-2024].

RAKOTOMAMONJY, A. Optimizing area under roc curve with svms. In: ROCAI. [S.l.: s.n.], 2004. p. 71–80.

RASAMOELINA, A. D.; ADJAILIA, F.; SINČÁK, P. A review of activation function for artificial neural network. In: IEEE. 2020 IEEE 18th World Symposium on Applied Machine Intelligence and Informatics (SAMI). [S.l.], 2020. p. 281–286.

RICARDO. Glaucoma É a principal causa de Cegueira irreversível no mundo; Veja Como Se Prevenir - SBG - Sociedade Brasileira de Glaucoma. 2024. Disponível em: <https://www.sbglaucoma.org.br/medico/glaucoma-e-a-principal-causa-de-cegueira-irreversivel-no-mundo-veja-como-se-prevenir/>.

RODRIGUES, G. B.; ABE, R. Y.; ZANGALLI, C.; SODRE, S. L.; DONINI, F. A.; COSTA, D. C.; LEITE, A.; FELIX, J. P.; TORIGOE, M.; DINIZ-FILHO, A. et al. Neovascular glaucoma: a review. International journal of retina and vitreous, Springer, v. 2, p. 1–10, 2016.

ROJAS, R.; ROJAS, R. The backpropagation algorithm. Neural networks: a systematic introduction, Springer, p. 149–182, 1996.

RUBY, U.; YENDAPALLI, V. et al. Binary cross entropy with deep learning technique for image classification. Int. J. Adv. Trends Comput. Sci. Eng, v. 9, n. 10, 2020.

SANDLER, M.; HOWARD, A.; ZHU, M.; ZHMOGINOV, A.; CHEN, L.-C. Mobilenetv2: Inverted residuals and linear bottlenecks. In: Proceedings of the IEEE conference on computer vision and pattern recognition. [S.l.: s.n.], 2018. p. 4510–4520.

SAXENA, A.; VYAS, A.; PARASHAR, L.; SINGH, U. A glaucoma detection using convolutional neural network. In: IEEE. 2020 international conference on electronics and sustainable communication systems (ICESC). [S.l.], 2020. p. 815–820.

SCHOPENHAUER, A. Parerga e Paralipomena Pequenos Escritos Filosóficos. [S.l.]: Lebooks Editora, 2021.

Scikit-learn. Ensembles: Gradient boosting, random forests, bagging, voting, stacking. [S.l.], 2025. User Guide, versão estável. Disponível em: <https://scikit-learn.org/stable/modules/ensemble.html>.

SENIOR, P. Estima-se que 64 milhões de pessoas tenham glaucoma no mundo. Será que você é uma delas? 2023. Disponível em: <https://www.preventsenior.com.br/blog/estima-se-que-64-milh%C3%B5es-de-pessoas-tenham-glaucoma-no-mundo-ser%C3%A1-que-voc%C3%AA-%C3%A9-uma-delas>.

SERTE, S.; SERENER, A. A generalized deep learning model for glaucoma detection. In: IEEE. 2019 3rd International symposium on multidisciplinary studies and innovative technologies (ISMSIT). [S.l.], 2019. p. 1–5.

SILVA, T. O pão que o viado amassou: contribuições da semiótica para o processamento de língua natural. Estudos Semióticos, v. 18, p. 70–92, 12 2022.

SIMIC, M. Hard vs. Soft Voting Classifiers. 2025. <https://www.baeldung.com/cs/hard-vs-soft-voting-classifiers>. Last updated: February 28, 2025; acesso em 19 de agosto de 2025.

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

SINGH, D.; SINGH, B. Investigating the impact of data normalization on classification performance. Applied Soft Computing, p. 105524, 05 2019.

STEINER, A.; KOLESNIKOV, A.; ZHAI, X.; WIGHTMAN, R.; USZKOREIT, J.; BEYER, L. How to train your ViT? Data, Augmentation, and Regularization in Vision Transformers. 2022. Disponível em: <https://arxiv.org/abs/2106.10270>.

TAJ, I. A.; SAJID, M.; KARIMOV, K. S. et al. An ensemble framework based on deep cnns architecture for glaucoma classification using fundus photography. Mathematical Biosciences and Engineering, AIMS Press, v. 18, n. 5, p. 5321–5347, 2021.

TAN, M.; LE, Q. EfficientNet: Rethinking model scaling for convolutional neural networks. In: CHAUDHURI, K.; SALAKHUTDINOV, R. (Ed.). Proceedings of the 36th International Conference on Machine Learning. PMLR, 2019. (Proceedings of Machine Learning Research, v. 97), p. 6105–6114. Disponível em: <https://proceedings.mlr.press/v97/tan19a.html>.

TAY, Y.; DEHGHANI, M.; BAHRI, D.; METZLER, D. Efficient Transformers: A Survey. 2022. Disponível em: <https://arxiv.org/abs/2009.06732>.

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.

TOHYE, T. G.; QIN, Z.; AL-ANTARI, M. A.; UKWUOMA, C. C.; LONSEKO, Z. M.; GU, Y. H. Ca-vit: Contour-guided and augmented vision transformers to enhance glaucoma classification using fundus images. Bioengineering, v. 11, n. 9, p. 887, 2024.

TU, Z.; TALEBI, H.; ZHANG, H.; YANG, F.; MILANFAR, P.; BOVIK, A.; LI, Y. Maxvit: Multi-axis vision transformer. In: SPRINGER. European conference on computer vision. [S.l.], 2022. p. 459–479.

U, R.; FUJITA, H.; BHANDARY, S.; GUDIGAR, A.; TAN, J. H.; ACHARYA, U. Deep convolution neural network for accurate diagnosis of glaucoma using digital fundus images. Information Sciences, v. 441, 02 2018.

Unimed. Glaucoma: prevenção e tratamento da doença. 2025. Acessado em: 09 mar. 2025. Disponível em: <https://viverbem.unimed.coop.br/saude-em-pauta/prevencao-e-tratamento-de-doencas/glauco-1/>.

UR;, C. Y. E. V. K. J. O. L. L. Automated detection of glaucoma using optical coherence tomography angiogram images. Computers in biology and medicine, U.S. National Library of Medicine, Oct 2019. Disponível em: <https://pubmed.ncbi.nlm.nih.gov/31698235/>.

VASWANI, A.; SHAZEER, N.; PARMAR, N.; USZKOREIT, J.; JONES, L.; GOMEZ, A. N.; KAISER, L.; POLOSUKHIN, I. Attention Is All You Need. 2017. Disponível em: <https://arxiv.org/abs/1706.03762>.

VELPULA, V. K.; SHARMA, D.; SHARMA, L. D.; ROY, A.; BHUYAN, M. K.; ALFARHOOD, S.; SAFRAN, M. Glaucoma detection with explainable ai using convolutional neural networks based feature extraction and machine learning classifiers. IET Image Processing, v. 18, n. 13, p. 3827–3853, 2024. Disponível em: <https://ietresearch.onlinelibrary.wiley.com/doi/abs/10.1049/ipr2.13211>.

VELPULA, V. K.; SHARMA, L. D. Multi-stage glaucoma classification using pre-trained convolutional neural networks and voting-based classifier fusion. Frontiers in Physiology, v. 14, 2023. ISSN 1664-042X. Disponível em: <https://www.frontiersin.org/journals/physiology/articles/10.3389/fphys.2023.1175881>.

WEINREB, R. N.; AUNG, T.; MEDEIROS, F. A. The pathophysiology and treatment of glaucoma: a review. Jama, American Medical Association, v. 311, n. 18, p. 1901–1911, 2014.

WU, J.; YU, S.; CHEN, W.; MA, K.; FU, R.; LIU, H.; DI, X.; ZHENG, Y. Leveraging Undiagnosed Data for Glaucoma Classification with Teacher-Student Learning. 2020. Disponível em: <https://arxiv.org/abs/2007.11355>.

YAKURA, H.; SHINOZAKI, S.; NISHIMURA, R.; OYAMA, Y.; SAKUMA, J. Malware analysis of imaged binary samples by convolutional neural network with attention mechanism. In: . [S.l.: s.n.], 2018. p. 127–134.

YANG, C.; QIAO, S.; YU, Q.; YUAN, X.; ZHU, Y.; YUILLE, A.; ADAM, H.; CHEN, L.-C. Moat: Alternating mobile convolution and attention brings strong vision models. arXiv preprint arXiv:2210.01820, 2022.

YURDAKUL, M.; UYAR, K.; TASDEMIR, S. Maxglavit: A novel lightweight vision transformer-based approach for early diagnosis of glaucoma stages from fundus images. arXiv preprint arXiv:2502.17154, 2025.

ZHANG, A.; LIPTON, Z. C.; LI, M.; SMOLA, A. J. Dive into Deep Learning. [S.l.]: Cambridge University Press, 2023. <https://D2L.ai>.

ZHANG, Z. Improved adam optimizer for deep neural networks. In: IEEE. 2018 IEEE/ACM 26th international symposium on quality of service (IWQoS). [S.l.], 2018. p. 1–2.

ZHANG, Z.; SABUNCU, M. Generalized cross entropy loss for training deep neural networks with noisy labels. In: BENGIO, S.; WALLACH, H.; LAROCHELLE, H.; GRAUMAN, K.; CESA-BIANCHI, N.; GARNETT, R. (Ed.). Advances in Neural Information Processing Systems. Curran Associates, Inc., 2018. v. 31. Disponível em: <https://proceedings.neurips.cc/paper_files/paper/2018/file/f2925f97bc13ad2852a7a551802feea0-Paper.pdf>.

ZHEN, Y.; WANG, L.; LIU, H.; ZHANG, J.; PU, J. Performance assessment of the deep learning technologies in grading glaucoma severity. arXiv preprint arXiv:1810.13376, 2018.

ZHOU, Z.-H. Machine learning. [S.l.]: Springer nature, 2021.

ZOU, F.; SHEN, L.; JIE, Z.; ZHANG, W.; LIU, W. A sufficient condition for convergences of adam and rmsprop. In: Proceedings of the IEEE/CVF Conference on computer vision and pattern recognition. [S.l.: s.n.], 2019. p. 11127–11135.