CNNs, Vision Transformers, and Ensembles for Glaucoma Classification: A Comparative Analysis

Arthur Celestino[1], **Guilherme Silva**[2], **Luan Pinto**[1],
Jadson Castro Getrudes[3], **Pedro Silva**[3]

[1] Postgraduate Program in Computer Science, Federal University of Ouro Preto
Ouro Preto – MG – Brazil
[2] Computing Department – Federal University of São João del-Rei
São João del-Rei – MG – Brasil
[3] Computing Department – Federal University of Ouro Preto
Ouro Preto – MG – Brazil
arthur.celestino@aluno.ufop.edu.br

Abstract. *Glaucoma is one of the leading causes of irreversible blindness worldwide, characterized by progressive damage to the optic nerve. Early diagnosis is essential to minimize the impact of the disease. In this work, neural architectures are explored to classify fundus images from the SMDG dataset, comparing pre-trained CNNs and transformers, as well as voting-based ensemble strategies. ResNet-50 and VGG-19 achieved consistent performance, but with lower sensitivity. Among the transformers, MaxViT achieved the highest recall, favoring positive-case detection, while ViT showed a balance between accuracy and sensitivity. Average voting outperformed individual models, reaching 89.45% accuracy.*

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. Neste trabalho, arquiteturas neurais são exploradas para classificar imagens de fundo de olho da base SMDG, comparando CNNs pré-treinadas e transformers, além de estratégias de ensemble por votação. ResNet-50 e VGG-19 obtiveram desempenho consistente, mas menor sensibilidade. Entre os transformers, MaxViT alcançou o maior recall, favorecendo a detecção de positivos, enquanto ViT apresentou equilíbrio entre acurácia e sensibilidade. A votação por média superou modelos individuais, atingindo 89,45% de acurácia.*

1. Introdução

O glaucoma é uma neuropatia óptica progressiva frequentemente associada ao aumento da pressão intraocular. Quando não diagnosticada e tratada adequadamente, a doença pode causar danos irreversíveis ao nervo óptico e evoluir para perda visual permanente ou cegueira [Cardoso 2024]. Trata-se de uma condição de alto impacto em saúde pública, especialmente em populações idosas [Tham et al. 2014]. Estima-se que a prevalência global de glaucoma de ângulo aberto deve aumentar de 80,5 milhões de casos em 2024 para 186,6 milhões até 2060, impulsionada pelo envelhecimento populacional e pelo aumento da miopia [Wang et al. 2026].

A detecção precoce é essencial, pois a progressão do glaucoma tende a ser lenta e inicialmente assintomática. Os sinais clínicos geralmente se manifestam em estágios mais

avançados, quando já pode haver perda progressiva da visão periférica e comprometimento funcional relevante [Prevent Senior 2023]. O diagnóstico é realizado por profissionais da saúde ocular a partir de exames clínicos, funcionais e de imagem [Wang et al. 2026]. Entre os procedimentos mais utilizados estão a tonometria, para medição da pressão intraocular; a avaliação do nervo óptico, frequentemente baseada na relação escavação/disco óptico; a gonioscopia, para análise do ângulo iridocorneano; a tomografia de coerência óptica (OCT), para avaliação estrutural das camadas retinianas; e o teste de campo visual, para mensuração de perdas funcionais [Chan et al. 2019]. Embora complementares, esses exames podem demandar infraestrutura especializada, tempo de execução e interpretação clínica, além de apresentarem limitações distintas de sensibilidade e precisão [Ashtari-Majlan et al. 2023].

Nesse contexto, métodos de inteligência artificial baseados em aprendizado profundo têm sido amplamente investigados como ferramentas de apoio à classificação de glaucoma em imagens de fundo de olho. Redes neurais convolucionais (CNNs) consolidaram-se como uma abordagem eficaz para extração automática de padrões visuais relevantes [Oh et al. 2021], sendo exploradas em diferentes configurações na literatura [Saxena et al. 2020; Chen et al. 2015; Velpula et al. 2024]. Estratégias de votação entre modelos também foram investigadas como forma de combinar previsões de arquiteturas distintas [Velpula and Sharma 2023]. Mais recentemente, modelos baseados em *transformers*, como Vision Transformers (ViTs), passaram a ser empregados nessa tarefa, embora usualmente apresentem maior complexidade computacional [Dosovitskiy et al. 2021; Yurdakul et al. 2025].

Este trabalho se posiciona como uma avaliação empírica de CNNs, *transformers* e estratégias de votação para classificação binária de glaucoma em imagens de fundo de olho. Em vez de propor uma nova arquitetura, o estudo investiga quando modelos individuais são suficientes e quando a combinação de modelos pode oferecer melhor equilíbrio entre desempenho global e sensibilidade clínica. A pergunta de pesquisa que orienta o trabalho é: "*a combinação de modelos heterogêneos por votação melhora o desempenho geral e o equilíbrio entre métricas clinicamente relevantes na classificação de glaucoma?*"

Os experimentos são conduzidos na base SMDG, considerando modelos convolucionais, modelos baseados em *transformers* e esquemas de votação. Os resultados indicam que CNNs clássicas podem alcançar desempenho global competitivo, mas apresentam maior limitação na identificação de casos positivos. Em contrapartida, modelos baseados em *transformers* aumentam a sensibilidade para a classe glaucoma, embora nem sempre apresentem o melhor desempenho agregado. A estratégia de *Soft Voting*, baseada na média das probabilidades preditivas, obteve o melhor balanço geral, alcançando 89,45% de acurácia e 0,9567 de área sob a curva ROC (AUC). Individualmente, a ResNet-50 obteve aproximadamente 87% de acurácia, enquanto o MaxViT alcançou 93% de *recall* para a classe positiva. Esses resultados sugerem que a combinação de previsões pode mitigar limitações específicas de modelos isolados, especialmente quando o objetivo é equilibrar sensibilidade clínica e desempenho global.

As contribuições deste trabalho são: **(i)** avaliação comparativa de CNNs clássicas e modelos baseados em *transformers* na classificação binária de glaucoma usando a SMDG-19; **(ii)** análise de estratégias de *ensemble*, comparando *Hard Voting* e *Soft Voting* para combinar modelos com comportamentos diferentes; e (iii) discussão do *trade-off* clínico

entre *recall* e precisão, mostrando que o MaxViT é mais sensível, enquanto o *Soft Voting* oferece o melhor equilíbrio geral.

O restante deste trabalho está organizado da seguinte forma. A Seção 2 apresenta trabalhos que inspiraram o atual trabalho. A Seção 3 apresenta a metodologia utilizada. A Seção 4 apresenta os experimentos realizados e os resultados obtidos e, por último, a Seção 5 apresenta conclusões e propostas de trabalhos futuros.

2. Trabalhos Relacionados

A classificação automática de glaucoma em imagens de fundo de olho tem sido amplamente investigada com modelos profundos de visão computacional, especialmente redes neurais convolucionais, arquiteturas baseadas em atenção e estratégias de fusão de modelos. Apesar dos avanços recentes, a tarefa permanece desafiadora devido à sutileza das alterações estruturais associadas aos estágios iniciais da doença, à variabilidade entre bases de dados e à escassez de conjuntos públicos extensos, bem anotados e clinicamente diversos [Rodrigues et al. 2016; Ashtari-Majlan et al. 2023]. Nesse cenário, torna-se relevante avaliar empiricamente não apenas o desempenho global dos modelos, mas também seu comportamento em métricas clinicamente críticas, como sensibilidade e especificidade.

Estudos baseados em redes neurais convolucionais mostram que essas arquiteturas continuam competitivas, especialmente quando associadas a pré-processamento, segmentação do disco óptico ou transferência de aprendizado. Trabalhos como [Zhen et al. 2018] e [Serte and Serener 2019] avaliaram arquiteturas pré-treinadas, incluindo VGG, ResNet, DenseNet, Inception, Xception e NASNetMobile, em múltiplas bases públicas, mas não superaram 80% de acurácia, indicando que a transferência direta de representações da ImageNet pode ser insuficiente para capturar padrões discriminativos de glaucoma. Em contraste, abordagens convolucionais mais específicas, como a rede de 18 camadas proposta por [U et al. 2018], alcançaram cerca de 98% de acurácia em uma base privada, embora com incertezas quanto à generalização. De forma semelhante, redes compactas e estratégias baseadas em recortes e média de predições, como em [Saxena et al. 2020] e [Chen et al. 2015], apresentaram resultados competitivos, mas dependem de etapas auxiliares de segmentação, recorte ou fusão, que podem afetar a reprodutibilidade e o custo de inferência.

Estratégias híbridas e de ensemble buscam aumentar a robustez ao combinar modelos ou representações complementares. Em Wu et al. [2020], uma abordagem teacher-student utilizou máscaras de segmentação do disco óptico e destilação de conhecimento, alcançando cerca de 95% de acurácia na base LAG, mas com maior complexidade de treinamento. Em Velpula et al. [2024], redes ResNet foram usadas como extratores de características combinados com classificadores como Support Vector Machine e k-nearest neighbors, obtendo 98,03% de acurácia no conjunto ACRIMA. Já Velpula and Sharma [2023] combinaram ResNet50, AlexNet, VGG19, DenseNet-201 e Inception-ResNet-v2 por votação majoritária em diferentes bases, com acurácias entre 99,57% na ACRIMA e 85,43% na HVD. Esses resultados indicam o potencial de ensembles, mas também evidenciam que seus ganhos devem ser analisados em relação à variabilidade entre bases e ao aumento proporcional do custo de inferência.

Mais recentemente, arquiteturas baseadas em transformers têm sido exploradas como alternativa às redes convolucionais. Em Hwang et al. [2023], Vision Transformers

foram comparados à ResNet-50 em seis bases públicas, incluindo Drishti-GS1, RIM-ONE DL, ORIGA, ACRIMA e REFUGE2. Os resultados indicaram que os Vision Transformers superaram as redes convolucionais em área sob a curva ROC, acurácia e F1-score em cinco das seis bases, especialmente em cenários com maior proporção de imagens negativas. Entretanto, também apresentaram maior tendência a falsos positivos em alguns conjuntos, reduzindo a especificidade. Outros trabalhos, como [Tohye et al. 2024] e Yurdakul et al. [2025], exploraram variações de Vision Transformers, Detection Transformer e MaxViT para classificação ou detecção de estruturas anatômicas relevantes, obtendo resultados promissores, mas frequentemente associados a maior demanda computacional, dependência de dados e desafios de generalização.

Em síntese, a literatura sugere que redes neurais convolucionais permanecem eficientes e competitivas, mas podem apresentar limitações em sensibilidade; transformers tendem a capturar relações globais e melhorar a detecção de casos positivos, embora possam reduzir a especificidade em alguns cenários; e ensembles oferecem um mecanismo simples para combinar modelos complementares, ao custo de maior inferência. Assim, este trabalho se diferencia ao realizar uma avaliação empírica comparativa entre métodos convolucionais, transformers e estratégias de votação para classificação de glaucoma em imagens de fundo de olho, com foco no *trade-off* entre sensibilidade clínica e desempenho global.

3. Configuração Experimental

Este trabalho realiza uma avaliação empírica de modelos convolucionais, modelos baseados em *transformers* e estratégias de votação para classificação binária de glaucoma em imagens de fundo de olho. O objetivo é comparar arquiteturas representativas sob um protocolo experimental comum, analisando o *trade-off* entre desempenho global e sensibilidade clínica. Em vez de propor uma nova arquitetura, o estudo investiga quando modelos isolados são suficientes e quando a combinação de predições por votação pode oferecer melhor equilíbrio entre identificação de casos positivos e desempenho geral.

Foi utilizada a base *Standardized Multi-Channel Dataset for Glaucoma* (SMDG), composta pela união de 19 conjuntos públicos frequentemente empregados na literatura de glaucoma. A base contém 7.499 imagens sem glaucoma, 4.817 imagens com glaucoma e 133 imagens rotuladas como suspeitas, conforme apresentado na Tabela 1. Por reunir imagens provenientes de diferentes fontes, a *Standardized Multi-Database Glaucoma* apresenta maior heterogeneidade que bases isoladas, tornando-se adequada para avaliar a robustez dos modelos em um cenário experimental mais diverso. As imagens foram previamente padronizadas por Kiefer [2023] por meio de um procedimento que remove regiões de fundo, centraliza a imagem de fundo de olho, preenche áreas ausentes e redimensiona as amostras para 512×512 pixels. Esse processo reduz variações de escala, enquadramento, posição e aparência introduzidas pela combinação de múltiplas bases, conforme proposto em [Kiefer et al. 2023].

O problema foi formulado como uma tarefa binária, na qual imagens com glaucoma foram consideradas classe positiva e imagens sem glaucoma foram consideradas classe negativa. As amostras rotuladas como suspeitas foram removidas devido à baixa representatividade no conjunto, correspondente a pouco mais de 1% do total, e por não contribuírem diretamente para o objetivo de distinguir casos positivos e negativos. Essa

Tabela 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**

exclusão constitui uma limitação do estudo, uma vez que casos ambíguos são particularmente relevantes em cenários reais de triagem, nos quais a distinção entre positivo, negativo e suspeito é clinicamente significativa. Também existe uma limitação no fato de que algumas bases que compõem a SMDG contribuem de forma desproporcional para uma das classes, como, por exemplo, a EyePACS-AIROGS, que contém exclusivamente imagens da classe glaucoma, o que pode introduzir vieses específicos de fonte e limitar a generalização dos resultados para populações e equipamentos de captura distintos. Após essa filtragem, os dados foram divididos em 70% para treinamento, 15% para validação e 15% para teste, seguindo uma configuração adotada em trabalhos recentes com a mesma base [Tohye et al. 2024]. A distribuição das classes foi monitorada em cada partição para evitar desbalanceamentos adicionais entre os subconjuntos.

O pré-processamento foi definido de acordo com os requisitos de entrada de cada família de modelos. Para as redes neurais convolucionais, as imagens foram redimensionadas para 124×124 pixels e normalizadas pela divisão dos valores de pixel por 255, restringindo-os ao intervalo $[0, 1]$. Essa resolução foi adotada para reduzir o custo computacional e permitir a comparação entre múltiplas arquiteturas sob as limitações de hardware disponíveis. Durante o treinamento das redes convolucionais, foi aplicado aumento de dados dinâmico com espelhamento horizontal e vertical, rotação aleatória de até $45°$, deslocamentos horizontal e vertical de até 20%, zoom de até 20% e preenchimento pelo método `nearest`. Para os modelos baseados em *transformers*, as imagens foram redimensionadas conforme a entrada esperada por cada arquitetura: 224×224 pixels para o Vision Transformer e 384×384 pixels para o MaxViT. No Vision Transformer, o processamento foi realizado pelo `ViTImageProcessor`; no MaxViT, as imagens foram normalizadas por canal com média $[0{,}5, 0{,}5, 0{,}5]$ e desvio padrão $[0{,}5, 0{,}5, 0{,}5]$. Não foi aplicado aumento de dados aos *transformers*, permitindo avaliar o desempenho dos modelos pré-treinados com mínima intervenção adicional, em linha com estudos que indicam bom desempenho de Vision Transformers pré-treinados mesmo com aumento de

dados reduzido em cenários de ajuste fino [Steiner et al. 2022].

Foram avaliadas três redes neurais convolucionais pré-treinadas na ImageNet: ResNet-50, VGG-19 e EfficientNet-B0. A ResNet-50 foi selecionada por suas conexões residuais, que facilitam o treinamento de redes profundas e são amplamente utilizadas em tarefas de visão computacional [Kundu 2023]. A VGG-19 foi incluída por sua recorrência em estudos de classificação de glaucoma e por sua arquitetura convolucional profunda baseada em filtros 3×3 [Simonyan and Zisserman 2014; Velpula and Sharma 2023; Zhen et al. 2018]. A EfficientNet-B0 foi adotada como alternativa mais eficiente, permitindo avaliar o balanço entre desempenho e custo computacional [Tan and Le 2019]. Em todos os casos, a base convolucional foi usada como extratora de características, seguida por uma camada de *Global Average Pooling*, duas camadas densas com 256 e 128 neurônios, ativação ReLU, normalização em lote, Dropout de 30% e regularização L2. A camada final contém um neurônio com ativação sigmoide, produzindo a probabilidade de a imagem pertencer à classe glaucoma. A predição binária foi obtida com limiar de 0,5.

Como representantes de arquiteturas baseadas em *transformers*, foram avaliados o Vision Transformer e o MaxViT. Para o Vision Transformer, foi utilizado o modelo pré-treinado *google/vit-base-patch16-224-in21k*, disponibilizado pela Hugging Face[1]. A cabeça de classificação original foi substituída por uma camada linear com duas unidades de saída, correspondentes às classes glaucoma e não glaucoma, preservando a estrutura interna do *encoder* e os pesos pré-treinados. O treinamento foi conduzido com a classe `Trainer` da biblioteca `transformers`, usando taxa de aprendizado de 1×10^{-5}, *batch size* igual a 16, 25 épocas, precisão mista com `fp16` e escalonador `cosine_with_restarts`. Para o MaxViT, foi utilizada a implementação `maxvit_base_tf_384.in1k`, disponível no repositório *timm*[2]. O modelo foi carregado com pesos pré-treinados na ImageNet-1K e teve sua camada de classificação substituída por uma camada linear binária. Diferentemente do Vision Transformer, o MaxViT foi treinado manualmente em PyTorch, com controle direto do ciclo de treinamento e avaliação.

Além dos modelos isolados, foram avaliadas duas estratégias de votação para combinar predições de arquiteturas complementares: *Soft Voting* e Hard Voting. A hipótese experimental é que redes neurais convolucionais e *transformers* podem cometer erros distintos, pois as primeiras tendem a enfatizar padrões locais, como bordas e texturas, enquanto os segundos podem capturar relações espaciais mais globais. No Soft Voting, as probabilidades preditas pelos K modelos são agregadas por média aritmética. Para uma imagem x, seja $p_i(x)$ a probabilidade atribuída pelo modelo i à classe glaucoma. A probabilidade agregada é definida por: $\bar{p}(x) = \frac{1}{K} \sum_{i=1}^{K} p_i(x)$.

A decisão final é obtida por limiarização: $\hat{y}_{\text{soft}} = \begin{cases} 1, & \text{se } \bar{p}(x) \geq 0,5 \\ 0, & \text{caso contrário} \end{cases}$.

No *Hard Voting*, cada probabilidade é primeiro convertida em uma predição binária com limiar de 0,5, e a classe final é determinada pela maioria dos votos:

$$\hat{y} * \text{hard} = \mathbb{I}\left(\frac{1}{K}\sum * i = 1^{K} \mathbb{I}(p_i(x) \geq 0{,}5) \geq 0{,}5\right), \tag{1}$$

[1]Disponível em: `https://huggingface.co/google/vit-base-patch16-224`. Acesso em junho de 2025.

[2]Disponível em: `https://huggingface.co/timm/maxvit_base_tf_384.in1k`. Acesso em julho de 2025.

em que $\mathbb{I}(\cdot)$ representa a função indicadora. Em caso de empate, a predição foi atribuída à classe não glaucoma. O *Soft Voting* preserva a informação probabilística dos modelos, enquanto o *Hard Voting* considera apenas a decisão discreta de cada classificador. Essa comparação permite avaliar se o uso das probabilidades melhora o equilíbrio entre sensibilidade clínica e desempenho global.

O desempenho foi avaliado por acurácia, precisão, revocação, F1-score e área sob a curva ROC. A acurácia foi mantida por sua ampla utilização em trabalhos relacionados, mas não foi tratada como métrica suficiente, pois pode mascarar falhas na identificação de casos positivos em bases desbalanceadas. A revocação foi considerada central para a análise de sensibilidade clínica, uma vez que mede a proporção de imagens com glaucoma corretamente identificadas. A precisão foi utilizada para quantificar a confiabilidade das predições positivas, enquanto o F1-score sintetiza o equilíbrio entre precisão e revocação. Por fim, a área sob a curva ROC foi reportada para avaliar a capacidade discriminativa dos modelos em diferentes limiares de decisão, especialmente em um cenário com distribuição desigual entre classes [Filho 2023; Hossin and Sulaiman 2015; Rakotomamonjy 2004].

4. Experimentos e Resultados

Esta seção apresenta a configuração experimental e os resultados obtidos na avaliação de modelos convolucionais, modelos baseados em *transformers* e estratégias de votação para classificação binária de glaucoma em imagens de fundo de olho. O objetivo é comparar o comportamento dessas abordagens sob o mesmo conjunto de teste, com foco no *trade-off* entre desempenho global e sensibilidade clínica. Em particular, avalia-se se modelos isolados são suficientes ou se a combinação de predições por votação melhora o equilíbrio entre revocação, precisão, F1-score e área sob a curva ROC.

Os experimentos foram executados em uma máquina com 16 GB de memória RAM, processador Intel Core i7-13650HX e GPU NVIDIA GeForce RTX 4060, em ambiente Windows 11. O código foi desenvolvido em Python 3.10.11. Para os modelos convolucionais, foram utilizadas TensorFlow 2.13, Keras 2.10, PyTorch 2.1.2+cu118, transformers 4.52.3 e huggingface-hub 0.32.1. As demais dependências estão disponíveis junto ao código-fonte. [3]

Para os modelos convolucionais, foi utilizado o otimizador RMSProp com taxa de aprendizado inicial de 1×10^{-4} e função de custo *Binary Cross-Entropy*. Para o Vision Transformer e o MaxViT, foi utilizado AdamW, com taxas de aprendizado de 1×10^{-5} e 1×10^{-4}, respectivamente, e função de custo *Categorical Cross-Entropy*. Todos os modelos foram treinados e avaliados com a mesma divisão fixa de treinamento, validação e teste. As redes ResNet-50, VGG-19 e EfficientNet-B0 foram treinadas por 50 épocas com *batch size* 16; o Vision Transformer foi treinado por 25 épocas com *batch size* 32; e o MaxViT foi treinado por 10 épocas com *batch size* 4. Devido ao custo computacional, os resultados correspondem a uma única execução experimental, o que deve ser considerado como limitação na interpretação estatística das diferenças entre modelos.

4.1. Resultados

A Tabela 2 consolida os resultados dos modelos convolucionais, dos modelos baseados em *transformers* e das estratégias de votação no conjunto de teste. Entre as redes convo-

[3] Disponível em: `https://github.com/ArthurCelestino/fundus_image_classification.git`.

lucionais, a ResNet-50 apresentou o melhor desempenho geral, com acurácia de 0,8723, precisão de 0,8904 e área sob a curva ROC de 0,9339, seguida pela VGG-19, que obteve resultados próximos. No entanto, ambas apresentaram revocação inferior à precisão, indicando menor capacidade de recuperar casos positivos. A EfficientNet-B0, embora tenha alcançado revocação de 0,8016, apresentou baixa precisão e menor desempenho global, sendo excluída das estratégias de votação posteriores. Esses resultados sugerem que redes convolucionais clássicas permanecem competitivas em precisão e separabilidade, mas ainda perdem uma fração relevante de casos de glaucoma.

Tabela 2. Resumo comparativo das métricas por modelo e estratégia de votação. SV = *Soft Voting*; HV = *Hard Voting*. Melhores resultados em negrito.

Modelo	Acurácia	Precisão	Revocação	F1-score	AUC
CNN EfficientNet-B0	0,6996	0,5769	0,8016	0,6710	0,8218
CNN ResNet50	0,8723	0,8904	0,7592	0,8196	0,9339
CNN VGG19	0,8604	0,8721	0,7436	0,8028	0,9257
Transformer Vit	0,8799	0,8559	0,8244	0,8398	0,9294
Transformer MaxVit	0,8415	0,7287	**0,9320**	0,8179	0,9528
SV CNN (ResNet50 + VGG19)	0,8750	0,9005	0,7564	0,8222	0,9413
HV CNN (ResNet50 + VGG19)	0,8620	**0,9295**	0,6912	0,7929	0,9413
SV *Transformers* (Vit + MaxVit)	0,8912	0,8561	0,8598	0,8580	**0,9567**
HV *Transformers* (Vit + MaxVit)	0,8880	0,8844	0,8130	0,8472	**0,9567**
SV CNN + *Transformers*	**0,8945**	0,8774	0,8414	**0,8590**	**0,9567**
HV CNN + *Transformers*	0,8874	0,9082	0,7847	0,8419	**0,9567**

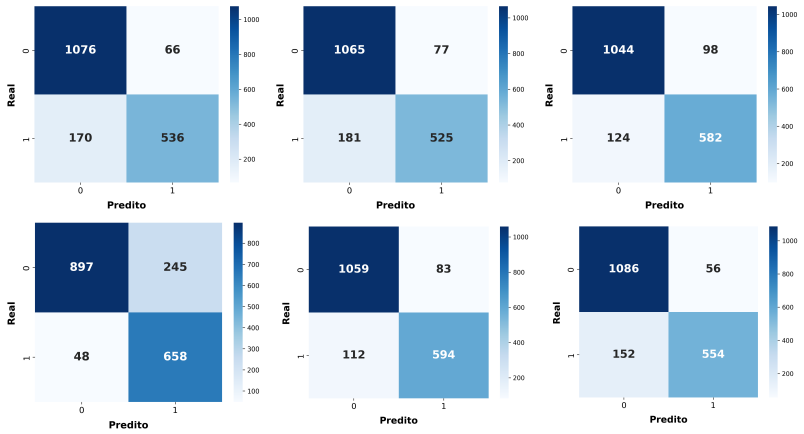
Os modelos baseados em *transformers* apresentam comportamento complementar. O Vision Transformer obteve desempenho mais equilibrado, com acurácia de 0,8799, precisão de 0,8559 e F1-score de 0,8398. O MaxViT, por outro lado, alcançou a maior revocação entre os modelos individuais (0,9320) e a maior área sob a curva ROC (0,9528), indicando maior sensibilidade para casos positivos, embora com redução de precisão. Assim, o MaxViT é mais adequado quando a prioridade é minimizar falsos negativos, enquanto o Vision Transformer oferece um melhor compromisso entre sensibilidade e controle de falsos positivos.

Entre as estratégias de fusão, observa-se que o *Soft Voting* apresentou melhor equilíbrio geral em todas as configurações avaliadas. Na fusão apenas entre CNNs (ResNet50 + VGG19), o *Soft Voting* superou o *Hard Voting* em acurácia, revocação e F1-score, enquanto o *Hard Voting* obteve a maior precisão geral (0,9295), indicando um comportamento mais conservador e menos sensível à classe positiva. Entre os modelos baseados em *Transformers* (ViT + MaxViT), os resultados foram superiores aos obtidos apenas com CNNs, especialmente em revocação e F1-score. O *Soft Voting* dos *Transformers* alcançou revocação de 0,8598 e F1-score de 0,8580, evidenciando melhor capacidade de identificação da classe positiva.

A fusão completa, combinando CNNs e *Transformers*, apresentou o melhor desempenho global. O *Soft Voting* atingiu a maior acurácia (0,8945) e o maior F1-score (0,8590), mantendo AUC de 0,9567. Esse resultado sugere que a agregação probabilística explora melhor a complementaridade entre arquiteturas distintas, enquanto o *Hard Voting* tende a priorizar a precisão em detrimento da revocação. Cabe notar, porém, que o ganho de acurácia do *Soft Voting* frente ao melhor modelo individual (89,45% vs. 87,99%) é pequeno e, dada a execução experimental única, deve ser interpretado com cautela quanto à sua robustez estatística.

As matrizes de confusão da Figura 1 reforçam os padrões observados nas métricas. ResNet-50 e VGG-19 preservam boa precisão, mas apresentam um maior número de falsos negativos. O MaxViT reduz esse tipo de erro, aumentando a sensibilidade, porém com maior incidência de falsos positivos. Já o *Soft Voting* (CNN + Transformers) apresenta uma distribuição mais equilibrada entre os erros, enquanto o *Hard Voting* (CNN + Transformers) favorece decisões mais conservadoras para a classe positiva.

Figura 1. Matrizes de confusão dos principais modelos e estratégias avaliados. Primeira linha ResNet-50, VGG-19 e Vision Transformer; segunda linha, MaxViT, *Soft Voting* e *Hard Voting*. Classe 0 = não glaucoma; 1 = glaucoma.



Além das métricas quantitativas, os mapas de ativação foram mantidos como análise qualitativa complementar, pois ajudam a interpretar os padrões de erro dos modelos. Na Figura 2, observa-se que, nos falsos negativos da ResNet-50, as ativações nem sempre coincidem com a região do disco óptico, onde se concentram sinais estruturais relevantes para o glaucoma. Nos falsos positivos das redes convolucionais, as ativações tendem a ocorrer em regiões mais próximas das estruturas anatômicas relevantes, mas ainda podem refletir padrões locais não discriminativos. Nos modelos baseados em *transformers*, os mapas sugerem ativações mais distribuídas: no MaxViT, falsos positivos apresentam ativação em múltiplas regiões da imagem, compatível com sua maior sensibilidade e maior tendência a classificar amostras ambíguas como positivas; no Vision Transformer, as ativações também são distribuídas, mas frequentemente incluem a região do disco óptico. Essa análise não deve ser interpretada como explicação causal dos erros, mas como evidência qualitativa consistente com o comportamento observado nas métricas.

A comparação com trabalhos que utilizaram a Standardized Multi-Database Glaucoma é apresentada na Tabela 3. Essa comparação deve ser interpretada como uma referência contextual, pois diferenças de particionamento, pré-processamento, filtragem de dados, sementes e estratégias de treinamento podem limitar comparações diretas. Ainda assim, observa-se que o *Soft Voting* (CNN + Transformers) obteve uma área sob a curva ROC competitiva, enquanto o MaxViT se destacou pela revocação, aproximando-se dos valores reportados por Tohye et al. [2024]. Cabe destacar que Tohye et al. [2024] avaliam seus modelos considerando as 5 classes originais da SMDG-19, enquanto este trabalho adota uma formulação binária (glaucoma vs. não glaucoma), o que limita a comparabilidade direta dos valores reportados. Em relação a Dosovitskiy et al. [2021], a comparação é limitada pela ausência de métricas completas, mas os valores da área sob a curva ROC obtidos neste trabalho são compatíveis.

Figura 2. Mapas de ativação para análise qualitativa dos erros. A primeira linha são os falsos positivos e a segunda linha, falsos negativos dos melhores modelos. A primeira coluna são do modelo ResNet-50, seguidos por Vision Transformer e MaxVit respectivametne.

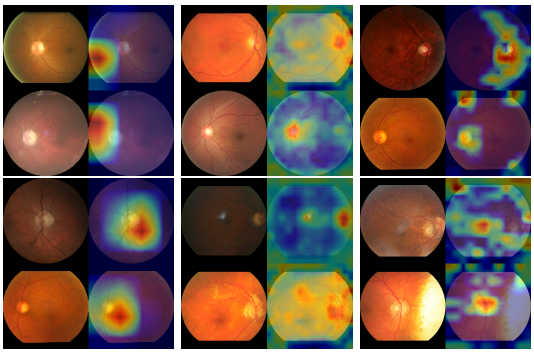


Tabela 3. Comparação com a literatura que utilizam a SMDG-19.

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 (CNN + Transformers)	**89,45**	87,74	84,14	**0,8590**	**0,9567**
Hard Voting (CNN + Transformers)	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	-

Em síntese, os resultados indicam que redes convolucionais clássicas permanecem competitivas, mas apresentam menor sensibilidade clínica; modelos baseados em *transformers*, especialmente o MaxViT, aumentam a recuperação de casos positivos; e o *Soft Voting* oferece o melhor balanço geral entre acurácia, F1-score e área sob a curva ROC. Assim, a escolha entre modelo isolado e ensemble depende do contexto de aplicação: sensibilidade máxima favorece o MaxViT, eficiência e equilíbrio favorecem o Vision Transformer, e desempenho global favorece o *Soft Voting*. A análise qualitativa por mapas de ativação pode ser mantida como material complementar ou apêndice, pois reforça a interpretação dos erros, mas não é essencial para a argumentação principal desta seção.

5. Conclusão

Este trabalho avaliou modelos convolucionais, *transformers* e estratégias de votação para classificação binária de glaucoma em imagens de fundo de olho na base Standardized Multi-Database Glaucoma. Os resultados indicam que ResNet-50 e VGG-19 mantêm boa acurácia e precisão, mas apresentam menor revocação, perdendo mais casos positivos. Já os *transformers* aumentaram a sensibilidade clínica: o MaxViT alcançou a maior revocação (93,20%), enquanto o Vision Transformer apresentou desempenho mais equilibrado. Entre as estratégias de fusão, o Soft Voting obteve o melhor balanço geral, com 89,45% de acurácia, 84,14% de revocação, F1-score de 0,8590 e área sob a curva ROC de 0,9567, indicando que a combinação probabilística explora a complementaridade entre CNNs e *transformers*. O Hard Voting foi mais preciso, mas reduziu a revocação ao descartar o grau de confiança das predições. Assim, a escolha do método depende do cenário: MaxViT favorece máxima sensibilidade, Vision Transformer oferece equilíbrio com menor custo, e Soft Voting maximiza o desempenho global. Como trabalhos futuros, recomenda-se investigar segmentação do disco óptico, calibração probabilística, ajuste de limiares, fusão ponderada, múltiplas sementes e custo inferencial, permitindo avaliar com maior robustez

o uso desses modelos em contextos de triagem assistida.

Agradecimentos e uso de IA

Os autores agradecem ao apoio da Fundação de Amparo à Pesquisa do Estado de Minas Gerais (FAPEMIG, projeto APQ-01768-24), ao Conselho Nacional de Desenvolvimento Científico e Tecnológico (CNPq), à Fundação Coordenação de Aperfeiçoamento de Pessoal de Nível Superior (CAPES), à Universidade Federal de Ouro Preto (PROPPI/UFOP) e a seu Programa de Pós-Graduação em Ciência da Computação (PPGCC/UFOP). Reconhecemos o uso de IA Generativa para auxiliar na revisão e refinamento deste artigo, que foi devidamente revisado, e assumimos total responsabilidade por seu conteúdo.

Referências

Ashtari-Majlan, M., Dehshibi, M. M., and Masip, D. (2023). Deep learning and computer vision for glaucoma detection: A review. *arXiv preprint arXiv:2307.16528*.

Cardoso, A. P. (2024). Glaucoma é a principal causa de cegueira irreversível no mundo; veja como se prevenir. `https://cultura.uol.com.br/noticias/66370_glaucoma-e-a-principal-causa-de-cegueira-irreversivel-no-mundo-veja-como-se-prevenir.html`. Acesso em: 11 ago. 2026.

Chan, Y. M., Ng, E. Y. K., Jahmunah, V., Koh, J. E. W., Lih, O. S., Leon, L. Y. W., and Acharya, U. R. (2019). Automated detection of glaucoma using optical coherence tomography angiogram images. *Computers in Biology and Medicine*, 115:103483.

Chen, X., Xu, Y., Yan, S., Wong, D. W. K., Wong, T. Y., and Liu, J. (2015). Automatic feature learning for glaucoma detection based on deep learning. In *Medical Image Computing and Computer-Assisted Intervention–MICCAI 2015: 18th International Conference, Munich, Germany, October 5-9, 2015, Proceedings, Part III 18*, pages 669–677. Springer.

Dosovitskiy, A., Beyer, L., Kolesnikov, A., Weissenborn, D., Zhai, X., Unterthiner, T., Dehghani, M., Minderer, M., Heigold, G., Gelly, S., Uszkoreit, J., and Houlsby, N. (2021). An image is worth 16x16 words: Transformers for image recognition at scale.

Filho, M. (2023). Precisão, recall e f1 score em machine learning. `https://towardsdatascience.com/a-look-at-precision-recall-and-f1-score-36b5fd0dd3ec`. Acesso em: 04 ago. 2023.

Hossin, M. and Sulaiman, M. N. (2015). A review on evaluation metrics for data classification evaluations. *International journal of data mining & knowledge management process*, 5(2):1.

Hwang, E. E., Chen, D., Han, Y., Jia, L., and Shan, J. (2023). Multi-dataset comparison of vision transformers and convolutional neural networks for detecting glaucomatous optic neuropathy from fundus photographs. *Bioengineering*, 10(11).

Kiefer, R. (2023). Smdg, a standardized fundus glaucoma dataset. `https://www.kaggle.com/ds/2329670`. DOI: 10.34740/KAGGLE/DS/2329670. Publicado por Kaggle.

Kiefer, R., Abid, M., Steen, J., Ardali, M. R., and Amjadian, E. (2023). 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*, ICISDM '23, page 24–31, New York, NY, USA. Association for Computing Machinery.

Kundu, N. (2023). Exploring resnet50: An in-depth look at the model architecture and code implementation.

Oh, S., Park, Y., Cho, K. J., and Kim, S. J. (2021). Explainable machine learning model for glaucoma diagnosis and its interpretation. *Diagnostics*, 11(3).

Prevent Senior (2023). Estima-se que 64 milhões de pessoas tenham glaucoma no mundo. será que você é uma delas? `https://tinyurl.com/39zkz992`. Publicado originalmente em preventsenior.com.br. Acesso em: 11 ago. 2026.

Rakotomamonjy, A. (2004). Optimizing area under roc curve with svms. In Hernández-Orallo, J., Ferri, C., Lachiche, N., and Flach, P. A., editors, *Proceedings of the 1st International Workshop on ROC Analysis in Artificial Intelligence (ROCAI-2004)*, pages 71–80, Valencia, Spain.

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. (2016). Neovascular glaucoma: a review. *International journal of retina and vitreous*, 2:1–10.

Saxena, A., Vyas, A., Parashar, L., and Singh, U. (2020). A glaucoma detection using convolutional neural network. In *2020 international conference on electronics and sustainable communication systems (ICESC)*, pages 815–820. IEEE.

Serte, S. and Serener, A. (2019). A generalized deep learning model for glaucoma detection. In *2019 3rd International symposium on multidisciplinary studies and innovative technologies (ISMSIT)*, pages 1–5. IEEE.

Simonyan, K. and Zisserman, A. (2014). Very deep convolutional networks for large-scale image recognition. *arXiv preprint arXiv:1409.1556*.

Steiner, A., Kolesnikov, A., Zhai, X., Wightman, R., Uszkoreit, J., and Beyer, L. (2022). How to train your vit? data, augmentation, and regularization in vision transformers.

Tan, M. and Le, Q. (2019). EfficientNet: Rethinking model scaling for convolutional neural networks. In Chaudhuri, K. and Salakhutdinov, R., editors, *Proceedings of the 36th International Conference on Machine Learning*, volume 97 of *Proceedings of Machine Learning Research*, pages 6105–6114. PMLR.

Tham, Y.-C., Li, X., Wong, T. Y., Quigley, H. A., Aung, T., and Cheng, C.-Y. (2014). Global prevalence of glaucoma and projections of glaucoma burden through 2040: a systematic review and meta-analysis. *Ophthalmology*, 121(11):2081–2090.

Tohye, T. G., Qin, Z., Al-Antari, M. A., Ukwuoma, C. C., Lonseko, Z. M., and Gu, Y. H. (2024). Ca-vit: Contour-guided and augmented vision transformers to enhance glaucoma classification using fundus images. *Bioengineering*, 11(9):887.

U, R., Fujita, H., Bhandary, S., Gudigar, A., Tan, J. H., and Acharya, U. (2018). Deep convolution neural network for accurate diagnosis of glaucoma using digital fundus images. *Information Sciences*, 441.

Velpula, V. K., Sharma, D., Sharma, L. D., Roy, A., Bhuyan, M. K., Alfarhood, S., and Safran, M. (2024). Glaucoma detection with explainable ai using convolutional neural networks based feature extraction and machine learning classifiers. *IET Image Processing*, 18(13):3827–3853.

Velpula, V. K. and Sharma, L. D. (2023). Multi-stage glaucoma classification using pre-trained convolutional neural networks and voting-based classifier fusion. *Frontiers in Physiology*, 14.

Wang, Z., Xue, C. C., Li, Y., Wu, Y., Pan, Z., Li, F., Cheung, C. Y., Ohno-Matsui, K., Friedman, D. S., Garway-Heath, D., et al. (2026). Global glaucoma prevalence: burden and projection to 2060. *American journal of ophthalmology*, 283:324–335.

Wu, J., Yu, S., Chen, W., Ma, K., Fu, R., Liu, H., Di, X., and Zheng, Y. (2020). Leveraging undiagnosed data for glaucoma classification with teacher-student learning.

Yurdakul, M., Uyar, K., and Tasdemir, S. (2025). Maxglavit: A novel lightweight vision transformer-based approach for early diagnosis of glaucoma stages from fundus images. *arXiv preprint arXiv:2502.17154*.

Zhen, Y., Wang, L., Liu, H., Zhang, J., and Pu, J. (2018). Performance assessment of the deep learning technologies in grading glaucoma severity. *arXiv preprint arXiv:1810.13376*.