Diagnóstico de Glaucoma em Retinografias de Oftalmoscópio Portátil Utilizando Ensemble Baseado em Transformers

Rodrigo Otávio C. Costa[1], **Patrik Oliveira Pimentel**[1], **Alexandre Cesar P. Pessoa**[1], **Geraldo Braz Júnior**[1], **João Dallyson S. Almeida**[1]

[1]Núcleo de Computação Aplicada – Universidade Federal do Maranhão (UFMA)
CEP 65085-580 – São Luís – MA – Brasil

{rodrigo.otavio, patrikop}@nca.ufma.br

{alexandre.pessoa, gebraz, jdallyson}@nca.ufma.br

Abstract. *This paper explores the use of Transformers-based models for the detection of glaucoma in portable ophthalmoscope retinograms, addressing the growing need for accessible diagnostics in regions with limited resources. Glaucoma is a chronic eye disease that can lead to irreversible blindness if not diagnosed early, and the implementation of automatic methodologies to aid screening is essential. The research uses a dataset made up of 2,000 fundus images, captured under practical conditions and with reduced quality, to train and evaluate the models. The results show that Transformers models, such as SwinV2, BEiT, DeiT and ViT, perform competitively compared to previous approaches based on convolutional networks, highlighting the potential of these models in classifying glaucoma from low-resolution retinograms. Combining the models' predictions in a Ensemble resulted in an average accuracy of 93.25%, demonstrating the effectiveness of the proposed approach.*

Resumo. *Este trabalho explora a utilização de modelos baseados em Transformers para a detecção de glaucoma em retinografias adquiridas por oftalmoscópio portátil, abordando a crescente necessidade de diagnósticos acessíveis em regiões com recursos limitados. O glaucoma é uma doença ocular crônica que pode levar à cegueira irreversível se não diagnosticada precocemente, sendo essencial a implementação de metodologias automáticas para auxiliar na triagem. A pesquisa utiliza um conjunto de dados composto por 2.000 imagens de fundo de olho, capturadas em condições práticas e com qualidade reduzida, para treinar e avaliar os modelos. Os resultados demonstram que os modelos de Transformers, como SwinV2, BEiT, DeiT e ViT, apresentam desempenho competitivo em comparação com abordagens anteriores baseadas em redes convolucionais, destacando o potencial desses modelos na classificação de glaucoma em retinografias de baixa resolução. A combinação das previsões dos modelos em um Ensemble resultou em uma acurácia média de 93,25%, evidenciando a eficácia da abordagem proposta.*

1. Introdução

O glaucoma é uma doença ocular crônica que afeta o funcionamento do nervo óptico, estrutura responsável por transmitir as informações visuais ao cérebro. O principal fator de

risco para a degeneração dessa estrutura é a elevação da pressão intraocular (PIO), decorrente de um desequilíbrio na produção e drenagem do humor aquoso. Nos estágios iniciais, a degeneração progressiva do nervo óptico geralmente passa despercebida, mas, com o tempo, pode levar à perda gradual da visão, tendo como início o campo visual periférico. Sem um tratamento adequado, o glaucoma pode evoluir para uma cegueira irreversível, logo, um diagnóstico precoce é essencial para evitar a progressão dessa doença [Jonas et al. 2017]. Uma das técnicas de diagnóstico envolve a análise manual do disco óptico por meio da oftalmoscopia, exame que permite avaliar alterações estruturais indicativas do glaucoma, como a relação entre o disco óptico e a escavação, onde um aumento anormal dessa relação pode indicar danos ao nervo óptico, auxiliando na detecção do glaucoma [Sarhan et al. 2019].

No mundo, estima-se que cerca de 2,2 bilhões de pessoas apresentem alguma deficiência visual, das quais pelo menos 1 bilhão poderia ser prevenida ou tratada, evidenciando a necessidade de tratamento em estágios iniciais em casos de doenças oculares [Organization et al. 2019]. Além disso, ainda que o número de oftalmologistas esteja crescendo em diversos países, um estudo global envolvendo 198 nações mostra uma distribuição desigual desses profissionais, resultando em um déficit significativo tanto no presente quanto nas projeções futuras, sobretudo em países de baixa e média renda [Resnikoff et al. 2020]. Sem contar a escassez de equipamentos oftalmológicos essenciais, que agrava ainda mais a situação, por exemplo, uma pesquisa nacional sobre a prática oftalmológica relacionada ao glaucoma na Nigéria revelou que 15-20% das clínicas não possuíam equipamentos básicos de diagnóstico para essa doença [Kyari et al. 2016].

Diante da desigualdade no acesso ao diagnóstico oftalmológico, metodologias automáticas de diagnóstico que possam auxiliar os especialistas nas triagens são fundamentais, principalmente em regiões carentes de profissionais e equipamentos especializados. Técnicas de visão computacional têm sido amplamente exploradas no âmbito da saúde para auxiliar na análise de imagens médicas, com redes neurais convolucionais (*Convolutional Neural Networks - CNNs*) demonstrando grande eficácia em tarefas como segmentação e classificação de estruturas oculares [Gulshan et al. 2016]. Recentemente, modelos baseados em *Transformers* vêm ganhando destaque nesse domínio [Takahashi et al. 2024]. Contudo, grande parte dos estudos para o diagnóstico do glaucoma depende de imagens de alta resolução obtidas por equipamentos avançados, limitando a aplicabilidade em contextos menos assistidos [Soofi et al. 2023].

Nesse cenário, a principal contribuição deste estudo está na proposição de uma abordagem de *Ensemble* baseada em modelos *Transformers* para a classificação do glaucoma em retinografias de baixa resolução, obtidas por oftalmoscópios portáteis acoplados a smartphones. Diferente da maioria dos estudos que priorizam imagens de alta qualidade e frequentemente utilizam *CNNs*, esta pesquisa explora o potencial dos Transformers para diagnósticos automatizados em cenários com recursos limitados. Ao combinar múltiplos modelos para aumentar a robustez e a confiabilidade da classificação, este trabalho busca não apenas melhorar a precisão do diagnóstico em imagens acessíveis, mas também ampliar a possibilidade de triagem em regiões carentes de infraestrutura especializada.

2. Trabalhos Relacionados

Esta seção apresenta trabalhos relacionados à classificação do glaucoma em retinografias, com destaque para abordagens que exploraram a viabilidade de metodologias aplicadas a imagens de menor qualidade. Para a seleção dos trabalhos analisados, foram considerados estudos que utilizaram bases de dados de resolução reduzida, similares à empregada neste trabalho, bem como pesquisas que exploraram imagens capturadas por dispositivos móveis, como smartphones.

Bragança et al. (2022) foram responsáveis pelo desenvolvimento de um conjunto de dados de imagens fundoscópicas com qualidade reduzida, utilizado para treinar e avaliar modelos de redes neurais convolucionais (*CNNs*). O estudo adotou um processo de treinamento baseado em validação cruzada *K-fold*. combinou modelos como DenseNet, MobileNet e InceptionResNet, formando um *Ensemble* para melhorar a classificação.

A pesquisa conduzida por Angara et al. (2024) propôs um *Ensemble* com VGG19, ResNet18 e DenseNet169 para aprimorar a generalização na classificação das imagens fundoscópicas, usando conjuntos de dados de alta e baixa resolução. Em Madhu et al. (2024) foram explorados modelos de CNNs pré-treinadas, como DenseNet-201 e Inception-ResNet-v2, otimizando a concatenação de características para melhorar a acurácia na classificação de imagens obtidas com oftalmoscópio portátil.

Mrad et al. (2022) propuseram um método rápido para triagem de glaucoma utilizando imagens capturadas por smartphones. Usando um classificador SVM, o estudo demonstrou alta precisão com qualidade moderada das imagens. Enquanto Moreira et al. (2021) exploraram o uso de redes de cápsulas (*Capsule Networks*) para a detecção de glaucoma, focando na preservação das hierarquias espaciais das imagens, minimizando as perdas de informação comuns em *CNNs* tradicionais. O estudo também enfatizou a importância de técnicas de pré-processamento, como remoção de ruído e segmentação vascular, para melhorar a acurácia da detecção.

Os trabalhos analisados demonstram a predominância do uso de redes neurais convolucionais (CNNs) para aprimorar a classificação de glaucoma. Algumas abordagens buscaram contornar essa limitação com pré-processamento e técnicas alternativas, como redes de cápsulas e classificação baseada em SVM. Diferente desses métodos, nosso trabalho propõe uma abordagem baseada em *Transformers*, explorando sua capacidade combinando diferentes modelos por meio de modelos *Ensemble*. A Tabela 1 apresenta um resumo comparativo das principais abordagens analisadas, destacando os modelos utilizados em cada estudo.

Tabela 1. Trabalhos relacionados à classificação de glaucoma.

Trabalho	Abordagem	Modelos Utilizados
[Bragança et al. 2022]	Ensemble de CNNs	DenseNet, MobileNet, InceptionResNet e outras *CNNs*
[Angara and Kim 2024]	Ensemble de CNNs	VGG19, ResNet18, DenseNet169
[Madhu et al. 2024]	CNNs pré-treinadas	DenseNet-201, InceptionResNet-v2
[Mrad et al. 2022]	Classificador SVM	SVM aplicado a imagens de smartphones
[Moreira et al. 2021]	Redes em Cápsulas	*Capsule Networks* treinadas em múltiplos datasets
Trabalho proposto	Ensemble de *Transformers*	Modelos baseados em *Transformers*

3. Fundamentação Teórica

Nesta seção, serão apresentadas as técnicas que fundamentam a metodologia utilizada no trabalho.

3.1. Modelos Baseados em Transformers

No presente trabalho, foram exploradas quatro arquiteturas baseadas em *Transformers*. Cada modelo possui abordagens específicas para otimizar o aprendizado em visão computacional, porém, todas compartilham de princípios fundamentais propostos por Vaswani et al. (2017): utilizam camadas de *self-attention* para modelar relações contextuais entre regiões da imagem, processam entradas visuais como sequências de *patches* (segmentos regulares da imagem), e empregam uma estrutura *encoder* com normalização de camadas e redes neurais *feed-forward*. Como os modelos analisados são voltados para tarefas de classificação e representação de imagens, eles utilizam apenas a parte do *encoder*, dispensando a necessidade do *decoder*. Embora existam variações incrementais entre os modelos, o núcleo operacional do *encoder* mantém-se alinhado ao padrão de atenção *multi-head*, conexões residuais e transformações lineares. A Figura 1 ilustra o bloco básico do *encoder* do *Transformer*, comum às arquiteturas empregadas neste trabalho.

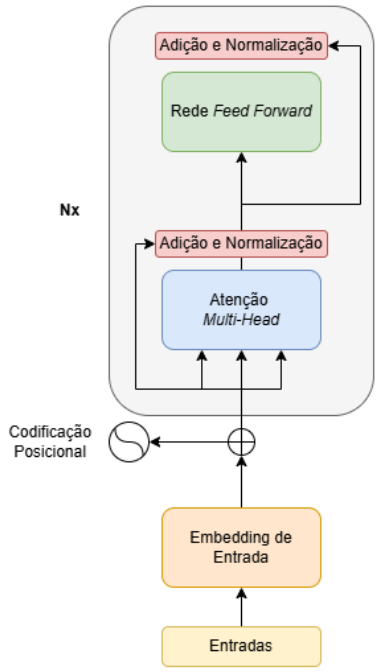

Figura 1. Estrutura do bloco *encoder* de um Transformer.
Adaptado de [Vaswani et al. 2017].

Dentre as arquiteturas exploradas neste trabalho, o *Vision Transformer* (ViT), introduzido por Dosovitskiy et al. (2020), apresentou a abordagem de dividir imagens em *patches* e tratá-los como sequência de *tokens*, aplicando a estrutura dos *Transformers* para capturar relações globais entre as regiões da imagem. O *Shifted Window Transformer* (Swin Transformer), proposto por Liu et al. (2021), aprimorou a eficiência dos *Transformers* para visão computacional ao introduzir a estrutura de janelas deslizantes, permitindo um processamento hierárquico que reduz o custo computacional e melhora a modelagem

de informações locais e globais. Posteriormente, o Swin Transformer V2 (SwinV2), expandiu essa abordagem, melhorando a escalabilidade para imagens de alta resolução e aprimorando a normalização dos parâmetros para maior estabilidade no treinamento [Liu et al. 2022].

Já o *Bidirectional Encoder Representation from Image Transformers* (BEiT), desenvolvido por Bao et al. (2021), adota estratégias de aprendizado auto-supervisionado inspiradas no Processamento de Linguagem Natural (PLN). Seu treinamento é baseado na predição de *patches* mascarados, permitindo uma melhor representação dos padrões visuais e tornando-o mais eficaz para tarefas de visão computacional. Por fim, o *Data-efficient Image Transformer* (DeiT), proposto por Touvron et al. (2021), foca na eficiência de dados, sendo projetado para treinamento com um menor volume de amostras e menos recursos computacionais. Sua principal inovação é o uso de um *token* de destilação, que auxilia o modelo a aprender de forma mais eficiente, tornando-o adequado para cenários com restrições de dados.

4. Metodologia

A metodologia proposta neste trabalho avalia a abordagem de *Transformers* para classificação de glaucoma em imagens de retinografia. Nesta seção, são apresentados os detalhes relacionados a base de imagens utilizada, os modelos empregados, o processo de treinamento e a construção dos modelos de *Ensemble*. A Figura 2 resume visualmente as etapas da metodologia adotada.

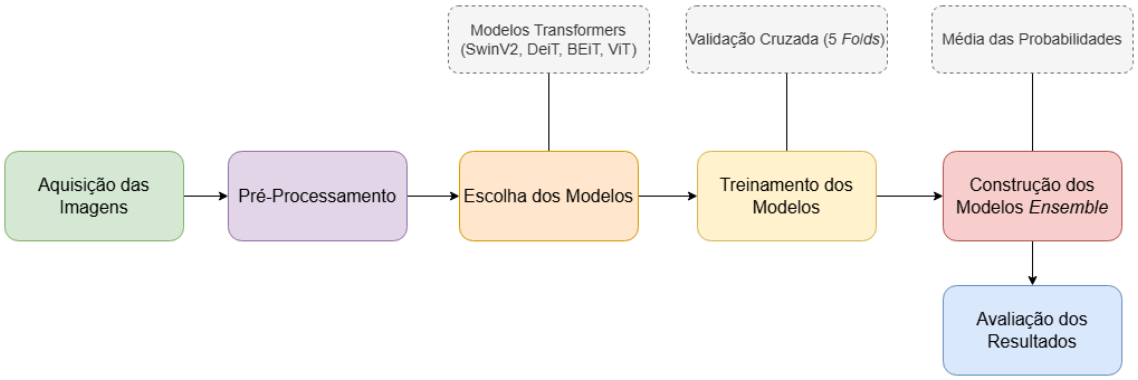

Figura 2. Etapas da metodologia proposta.

4.1. Base de Imagens

O conjunto de imagens utilizado neste estudo é o *Brazil Glaucoma* (BrG dataset), composto por 2.000 imagens de fundo de olho obtidas de 1.000 voluntários, sendo 500 pacientes diagnosticados com glaucoma e 500 indivíduos que não apresentam a doença. As imagens foram capturadas em duas unidades de saúde no Brasil, localizadas em Minas Gerais. A aquisição das imagens foi realizada sem dilatação pupilar e por profissionais não médicos, proporcionando um cenário mais próximo da prática clínica em ambientes com recursos limitados [Bragança et al. 2022].

Além disso, as imagens do *dataset* apresentam qualidade reduzida em comparação a bases de dados tradicionais. A captura foi realizada utilizando um oftalmoscópio panóptico acoplado a um aparelho celular, permitindo a obtenção de imagens de forma portátil e acessível. A Figura 3 ilustra exemplos das imagens capturadas para o *BrG dataset*.

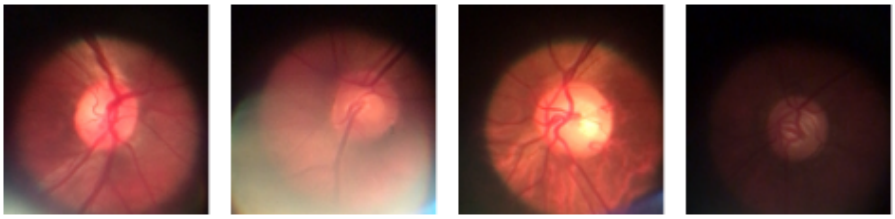

Figura 3. Amostras do Brazil Glaucoma Dataset.

4.2. Modelos Utilizados e Preparação das Imagens

Neste trabalho, foram utilizados modelos baseados em *Transformers*, mais especificamente: ViT, SwinV2, DeiT e BEiT. Todos os modelos estão disponíveis na biblioteca Hugging Face Transformers e apresentam versões pré-treinadas no *ImageNet*, um dos maiores e mais amplamente utilizado conjunto de dados para aprendizado em visão computacional [Wolf et al. 2020, Deng et al. 2009]. Os modelos foram utilizados para a tarefa de classificação das imagens de fundo de olho, distinguindo entre imagens com e sem glaucoma. A Tabela 2 mostra os modelos utilizados, seus respectivos tamanhos de entrada e os conjuntos de dados nos quais foram pré-treinados, destacando que alguns foram pré-treinados no *ImageNet-1K*, que contém aproximadamente 1,2 milhão de imagens classificadas em 1.000 classes, enquanto outros utilizaram *ImageNet-21K*, uma versão expandida com cerca de 14 milhões de imagens distribuídas em mais de 21.000 classes.

Tabela 2. Modelos *Transformers* pré-treinados utilizados neste estudo.

Modelo	Tamanho de Entrada	Pré-treinado em
SwinV2 [Liu et al. 2022]	256 × 256	ImageNet-1K
BEiT [Bao et al. 2021]	224 × 224	ImageNet-21K
DeiT [Touvron et al. 2021]	224 × 224	ImageNet-1K
ViT [Wu et al. 2020]	224 × 224	ImageNet-21K

O pipeline de treinamento foi padronizado para todos os modelos, alterando apenas a arquitetura utilizada e seu respectivo processador de imagens, garantindo assim que as configurações de treinamento e hiperparâmetros fossem idênticas em todos os experimentos. Ademais, essa abordagem assegura reprodutibilidade e comparabilidade entre os modelos.

As imagens foram pré-processadas utilizando transformações comuns em redes neurais, incluindo o redimensionamento para o tamanho exigido pelo modelo, *data augmentation* por meio de inversão horizontal e rotações, e, por fim, a normalização, que é realizada conforme os padrões do modelo pré-treinado.

4.3. Treinamento dos Modelos

Os pesos dos modelos foram inicializados com os pesos pré-treinados no ImageNet, conforme descrito na seção anterior. Estabelecemos o número de épocas em 10 e o tamanho do *batch* em 32. A técnica de treinamento utilizada foi validação cruzada *k-fold* com 5 *folds*. Nesse método, o conjunto de dados é dividido em 5 subconjuntos (*folds*), e o treinamento é repetido 5 vezes. A cada iteração, um dos *folds* é utilizado para validação, enquanto os demais são empregados no treinamento. Dessa forma, todos os dados são

utilizados tanto para treino quanto para validação ao longo do processo, reduzindo a dependência de uma única divisão entre treino e teste.

Ao final do treinamento de cada *fold*, foram obtidas as predições para as imagens do respectivo conjunto de validação. Ao final dos cinco *folds*, essas predições foram agrupadas para calcular as métricas gerais de desempenho, obtendo-se a média e o desvio padrão entre os *folds*. Esse procedimento garantiu uma avaliação mais consistente e robusta dos modelos testados.

4.4. Construção dos Modelos de Ensemble

A técnica de *Ensemble* combina previsões de múltiplos modelos para melhorar o desempenho preditivo, explorando os pontos fortes de cada um e aumentando a robustez das predições. Diferentes abordagens podem ser utilizadas para a agregação dos resultados, como votação majoritária e média das probabilidades. Neste trabalho, optamos pela estratégia de média das probabilidades, pois essa abordagem considera o nível de confiança atribuído por cada modelo às suas respectivas predições, em vez de apenas considerar a classe mais votada.

Após o treinamento individual dos modelos, os resultados foram combinados para formar um *Ensemble*. Para cada *fold* da validação cruzada, as probabilidades de cada classe (normal ou glaucoma) fornecidas pelos quatro modelos foram extraídas e, em seguida, calculamos a média dessas probabilidades para cada imagem. A classe com maior probabilidade média foi atribuída como a predição final do *Ensemble*.

Esse procedimento resultou na criação de cinco modelos distintos, correspondentes aos cinco *folds* da validação cruzada, garantindo que todas as amostras fossem avaliadas de maneira consistente em todos os modelos. A Figura 4 ilustra o processo de composição do *Ensemble*, nesse caso, para uma entrada específica.

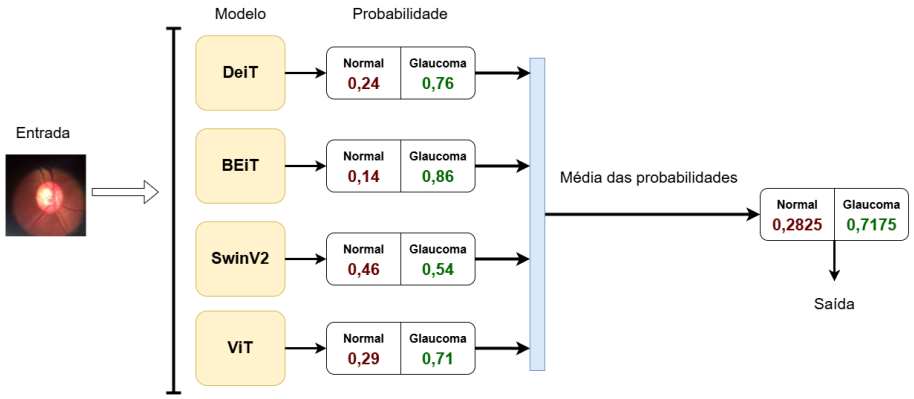

Figura 4. Etapas do Ensemble.

5. Resultados

Nesta seção, apresentamos os resultados obtidos com os modelos individuais e os modelos de *Ensemble*, destacando métricas relevantes para avaliar o desempenho dos métodos na classificação das imagens de fundo de olho.

Inicialmente, apresentamos as métricas utilizadas para a avaliação dos modelos propostos:

Acurácia ($Accuracy$**)**: Mede a proporção de imagens corretamente classificadas sobre o total de imagens avaliadas.

$$ACC = \frac{TP + TN}{TP + TN + FP + FN} \tag{1}$$

Especificidade ($Specificity$**)**: Mede a proporção de imagens normais (sem glaucoma) corretamente classificadas pelo modelo, sendo uma métrica relevante para evitar falsos diagnósticos positivos.

$$Specificity = \frac{TN}{TN + FP} \tag{2}$$

Precisão ($Precision$**)**: Representa a fração de imagens classificadas como glaucoma que realmente possuem a doença.

$$Precision = \frac{TP}{TP + FP} \tag{3}$$

Sensibilidade ($Recall$**)**: Mede a proporção de imagens com glaucoma que foram corretamente identificadas pelo modelo.

$$Recall = \frac{TP}{TP + FN} \tag{4}$$

F1-score ($F1$**)**: Combina precisão e sensibilidade em uma única métrica, sendo especialmente útil em conjuntos de dados desbalanceados, onde o número de imagens com e sem glaucoma pode ser desigual.

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

Área sob a curva ROC (AUC): A AUC (*Area Under the Curve*) avalia a capacidade do modelo de distinguir entre imagens com e sem glaucoma. Calculada a partir da curva ROC (*Receiver Operating Characteristic*), que relaciona sensibilidade em função da taxa de falsos positivos para diferentes limiares de decisão. Um valor próximo a 1,0 indica alta capacidade discriminativa, sendo especialmente útil para conjuntos de dados desbalanceados entre imagens saudáveis e com glaucoma.

O modelo foi avaliado utilizando a validação cruzada, garantindo que todas as imagens fossem utilizadas no treinamento e na validação. A Tabela 3 apresenta as métricas médias e seus respectivos desvios padrão para cada modelo.

Tabela 3. Desempenho médio dos modelos na validação cruzada.

Modelo	Acurácia	Especificidade	Precisão	Sensibilidade	F1	AUC
SwinV2	91,90% ± 1,93%	92,60% ± 3,27%	92,56% ± 2,64%	91,20% ± 2,32%	91,85% ± 1,91%	97,29% ± 0,95%
BEiT	92,40% ± 1,29%	92,80% ± 1,40%	92,74% ± 1,26%	92,00% ± 1,48%	92,37% ± 1,30%	97,49% ± 1,02%
DeiT	93,25% ± 1,76%	93,10% ± 3,93%	93,26% ± 3,09%	93,40% ± 2,89%	93,26% ± 1,73%	97,64% ± 0,64%
ViT	92,10% ± 1,53%	90,20% ± 3,19%	90,63% ± 2,49%	94,00% ± 1,10%	92,26% ± 1,41%	97,01% ± 1,01%

Os resultados indicam que o DeiT foi o modelo mais robusto, conseguindo um bom equilíbrio entre precisão e sensibilidade, enquanto o ViT mostrou maior sensibilidade à classe positiva (glaucoma), útil em cenários que minimizar falsos negativos é crucial. A SwinV2, apesar de ter valores competitivos, apresentou a menor acurácia e o menor F1-score, enquanto o BEiT apresentou um desempenho equilibrado, ficando ligeiramente inferior ao DeiT. A utilização da validação cruzada com 5 *folds* foi essencial para avaliar a capacidade de generalização dos modelos, pois a técnica reduz o risco de sobreajuste e fornece uma estimativa mais confiável do desempenho em dados não vistos, especialmente em aplicações médicas como o diagnóstico de glaucoma [Wilimitis and Walsh 2023].

Após a avaliação individual dos modelos, suas predições foram combinadas para construir os modelos de *Ensemble*, seguindo o método descrito na Seção 4.4. A Tabela 4 apresenta o desempenho desses modelos, detalhando as métricas obtidas em cada um dos cinco *folds*, além da média geral das métricas, acompanhada do desvio padrão, permitindo avaliar a consistência e a robustez dos resultados.

Tabela 4. Métricas obtidas no Ensemble por fold.

Fold	Acurácia	Especificidade	Precisão	Sensibilidade	F1	AUC
Fold 1	95,25%	95,50%	95,48%	95,00%	95,24%	99,02%
Fold 2	92,00%	89,50%	90,00%	94,50%	92,20%	97,40%
Fold 3	95,00%	95,50%	95,45%	94,50%	94,97%	98,38%
Fold 4	92,50%	94,00%	93,81%	91,00%	92,39%	98,05%
Fold 5	91,50%	92,50%	92,35%	90,50%	91,41%	96,91%
Média	93,25% $\pm$ 1,57%	93,40% $\pm$ 2,24%	93,42% $\pm$ 2,07%	93,10% $\pm$ 1,93%	93,24% $\pm$ 1,56%	97,95% $\pm$ 0,74%

Os valores de métrica alcançados assinalam que o *Ensemble* superou a acurácia dos modelos individuais, indicando que a combinação de predições reduziu erros gerais do modelo. Os valores médios de precisão (93,42%), sensibilidade (93,10%) e especificidade (93,40%) demonstram uma estabilidade e consistência do modelo, sem viés para nenhuma classe, enquanto a métrica ROC-AUC (97,95%) confirma alta capacidade de discriminação entre imagens normais e glaucomatosas. No mais, os baixos desvios padrão refletem a consistência do *Ensemble* entre os diferentes *folds*.

A Tabela 5 compara o desempenho do *Ensemble* com estudos anteriores que utilizaram o mesmo conjunto de imagens. Angara et al. (2024) alcançaram acurácia de 93,85%, enquanto Bragança et al. (2022) obtiveram 90,5%. A acurácia média de 93,25% do trabalho proposto evidencia que a abordagem alcança resultados equiparáveis ao estado da arte, mesmo utilizando somente um conjunto de dados para treinamento, diferentemente de Angara et al. (2024), que empregaram dois.

Tabela 5. Comparação das métricas do Ensemble com trabalhos relacionados.

Modelo	Acurácia	Especificidade	Precisão	Sensibilidade	F1	AUC
Bragança et al. (2022)	90,85%	96,00%	95,50%	85,00%	89,90%	96,50%
Angara et al. (2024)	93,85%	93,05%	-	94,67%	-	98,41%
Abordagem proposta	93,25% $\pm$ 1,57%	93,40% $\pm$ 2,24%	93,42% $\pm$ 2,07%	93,10% $\pm$ 1,93%	93,24% $\pm$ 1,56%	97,95% $\pm$ 0,74%

Diferentemente de abordagens anteriores baseadas em CNNs, como VGG19, ResNet18 e DenseNet, este trabalho utiliza arquiteturas baseadas em *Transformers*, que se destacam pela captura de relações globais via *self-attetion*. Essa característica é vantajosa em retinografias de baixa resolução, nas quais padrões de glaucoma, como alterações no

disco óptico, podem não ser locais. Comparadas às CNNs, que dependem de convoluções locais, os *Transformers* oferecem flexibilidade em imagens de qualidade variável, resultando em um bom desempenho neste contexto.

A Figura 5 apresenta dois exemplos de imagens do conjunto de dados, obtidas por oftalmoscópio portátil — característica que costuma ser um desafio adicional para a detecção precisa de glaucoma —, analisadas pelo *Ensemble*. À esquerda (Figura 5-A), observa-se um caso em que o modelo falhou ao tentar identificar o glaucoma. Neste exemplo, a baixa qualidade e o contraste reduzido entre as cores da retina e do disco óptico dificultaram a extração das características relevantes, comprometendo a capacidade do modelo de realizar uma identificação precisa da patologia. Em contrapartida, na imagem à direita (Figura 5-B), o modelo obteve sucesso ao identificar corretamente o glaucoma, mesmo diante das limitações visuais presentes, como iluminação irregular e artefatos luminosos.

Esses resultados ressaltam não apenas a eficiência da abordagem baseada em *transformers*, mas também destacam seu potencial significativo para lidar com a classificação de glaucoma em cenários desafiadores, abrindo espaço para futuras investigações que explorem diferentes bases de dados, técnicas de otimização das arquiteturas e ajustes específicos para esse domínio.

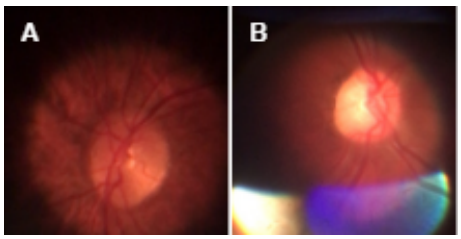

Figura 5. Estudo de casos de falha (A) e sucesso (B) da classificação do *Ensemble*.

6. Conclusão

Este trabalho apresentou um método baseado em modelos *Transformers* para a detecção de glaucoma em imagens de fundo de olho, utilizando modelos pré-treinados e modelos *Ensemble* para melhorar a robustez das predições. Os resultados demonstram que essa abordagem alcança um desempenho comparável ao de redes convolucionais, indicando o potencial dos *Transformers* para a análise de retinografias, inclusive em cenários com imagens de oftalmoscópio portátil.

Apesar dos resultados promissores, há oportunidades para aprimoramento. Como trabalho futuro, a aplicação de técnicas de processamento de imagem, como redução de ruído com filtros bilaterais e realce de contraste com *Contrast Limited Adaptive Histogram Equalization* (CLAHE), pode melhorar a qualidade visual das retinografias, tornando as características do glaucoma mais evidentes para os modelos. Além disso, o aumento de dados com bases externas pode melhorar a generalização, reduzindo a dependência de um único conjunto de imagens, possibilitando uma aplicabilidade maior em outros contextos clínicos. Por fim, ajustes mais refinados na configuração dos modelos, como escolha de hiperparâmetros, técnicas de regularização e otimização, podem levar a um desempenho superior.

7. Agradecimentos

Os autores agradecem o apoio financeiro da Coordenação de Aperfeiçoamento de Pessoal de Nível Superior (CAPES) - Finance Code 001, Conselho Nacional de Desenvolvimento Científico e Tecnológico (CNPq), e Fundação de Amparo à Pesquisa e ao Desenvolvimento Científico e Tecnológico do Maranhão (FAPEMA).

Referências

Angara, S. and Kim, J. (2024). Deep ensemble learning for classification of glaucoma from smartphone fundus images. In *2024 IEEE 37th International Symposium on Computer-Based Medical Systems (CBMS)*, pages 412–417. IEEE.

Bao, H., Dong, L., Piao, S., and Wei, F. (2021). Beit: Bert pre-training of image transformers. *arXiv preprint arXiv:2106.08254*.

Bragança, C. P., Torres, J. M., Soares, C. P. d. A., and Macedo, L. O. (2022). Detection of glaucoma on fundus images using deep learning on a new image set obtained with a smartphone and handheld ophthalmoscope. In *Healthcare*, volume 10, page 2345. MDPI.

Deng, J., Dong, W., Socher, R., Li, L.-J., Li, K., and Fei-Fei, L. (2009). Imagenet: A large-scale hierarchical image database. In *2009 IEEE conference on computer vision and pattern recognition*, pages 248–255. Ieee.

Dosovitskiy, A., Beyer, L., Kolesnikov, A., Weissenborn, D., Zhai, X., Unterthiner, T., Dehghani, M., Minderer, M., Heigold, G., Gelly, S., et al. (2020). An image is worth 16x16 words: Transformers for image recognition at scale. *arXiv preprint arXiv:2010.11929*.

Gulshan, V., Peng, L., Coram, M., Stumpe, M. C., Wu, D., Narayanaswamy, A., Venugopalan, S., Widner, K., Madams, T., Cuadros, J., et al. (2016). Development and validation of a deep learning algorithm for detection of diabetic retinopathy in retinal fundus photographs. *jama*, 316(22):2402–2410.

Jonas, J. B., Aung, T., Bourne, R. R., Bron, A. M., Ritch, R., and Panda-Jonas, S. (2017). Glaucoma. *The Lancet*, 390(10108):2183–2193.

Kyari, F., Nolan, W., and Gilbert, C. (2016). Ophthalmologists' practice patterns and challenges in achieving optimal management for glaucoma in nigeria: results from a nationwide survey. *BMJ open*, 6(10):e012230.

Liu, Z., Hu, H., Lin, Y., Yao, Z., Xie, Z., Wei, Y., Ning, J., Cao, Y., Zhang, Z., Dong, L., et al. (2022). Swin transformer v2: Scaling up capacity and resolution. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 12009–12019.

Liu, Z., Lin, Y., Cao, Y., Hu, H., Wei, Y., Zhang, Z., Lin, S., and Guo, B. (2021). Swin transformer: Hierarchical vision transformer using shifted windows. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 10012–10022.

Madhu, K., John, S. M., Joseph, A., and Abraham, B. (2024). Glaucoma diagnosis from smartphone captured fundus images using deep learning. In *2024 11th International Conference on Advances in Computing and Communications (ICACC)*, pages 1–6. IEEE.

Moreira, J. M. M., de Almeida, J. D. S., Junior, G. B., and de Paiva, A. C. (2021). Detecçao de glaucoma usando redes em cápsula. In *Simpósio Brasileiro de Computação Aplicada à Saúde (SBCAS)*, pages 188–199. SBC.

Mrad, Y., Elloumi, Y., Akil, M., and Bedoui, M. H. (2022). A fast and accurate method for glaucoma screening from smartphone-captured fundus images. *Irbm*, 43(4):279–289.

Organization, W. H. et al. (2019). World report on vision. In *World report on vision*.

Resnikoff, S., Lansingh, V. C., Washburn, L., Felch, W., Gauthier, T.-M., Taylor, H. R., Eckert, K., Parke, D., and Wiedemann, P. (2020). Estimated number of ophthalmologists worldwide (international council of ophthalmology update): will we meet the needs? *British Journal of Ophthalmology*, 104(4):588–592.

Sarhan, A., Rokne, J., and Alhajj, R. (2019). Glaucoma detection using image processing techniques: A literature review. *Computerized Medical Imaging and Graphics*, 78:101657.

Soofi, A. A. et al. (2023). Exploring deep learning techniques for glaucoma detection: a comprehensive review. *arXiv preprint arXiv:2311.01425*.

Takahashi, S., Sakaguchi, Y., Kouno, N., Takasawa, K., Ishizu, K., Akagi, Y., Aoyama, R., Teraya, N., Bolatkan, A., Shinkai, N., et al. (2024). Comparison of vision transformers and convolutional neural networks in medical image analysis: a systematic review. *Journal of Medical Systems*, 48(1):84.

Touvron, H., Cord, M., Douze, M., Massa, F., Sablayrolles, A., and Jégou, H. (2021). Training data-efficient image transformers & distillation through attention. In *International conference on machine learning*, pages 10347–10357. PMLR.

Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, Ł., and Polosukhin, I. (2017). Attention is all you need. *Advances in neural information processing systems*, 30.

Wilimitis, D. and Walsh, C. G. (2023). Practical considerations and applied examples of cross-validation for model development and evaluation in health care: tutorial. *Jmir ai*, 2:e49023.

Wolf, T., Debut, L., Sanh, V., Chaumond, J., Delangue, C., Moi, A., Cistac, P., Rault, T., Louf, R., Funtowicz, M., Davison, J., Shleifer, S., von Platen, P., Ma, C., Jernite, Y., Plu, J., Xu, C., Scao, T. L., Gugger, S., Drame, M., Lhoest, Q., and Rush, A. M. (2020). Transformers: State-of-the-art natural language processing. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, pages 38–45, Online. Association for Computational Linguistics.

Wu, B., Xu, C., Dai, X., Wan, A., Zhang, P., Yan, Z., Tomizuka, M., Gonzalez, J., Keutzer, K., and Vajda, P. (2020). Visual transformers: Token-based image representation and processing for computer vision.