Comparação de arquiteturas CNNs e transformers na triagem automatizada de retinopatia diabética

Danilo Leite[1], Roberto Mendes[2], Arthur Custódio[2], Alline Veloso[2], Sabrina Ferraz[2], Mateus Ramalho[3], José Câmara[1], Ronei Moraes[4]

[1]Universidade de Trás-os-Montes e Alto Douro (UTAD)
Vila Real, Portugal

[2]Programa de Pós-graduação em Modelos de Decisão e Saúde
Universidade Federal da Paraíba (UFPB)
João Pessoa – PB – Brasil

[3]Centro Universitário de João Pessoa (Unipê)
João Pessoa – PB – Brasil

[4]Departamento de Estatística
Universidade Federal da Paraíba (UFPB)
João Pessoa – PB – Brasil

danilol@utad.pt, roberto.mendes2@academico.ufpb.br, arthur.custodio2@academico.ufpb.br, lilionvjp@gmail.com, sabrina.ferraz@academico.ufpb.br, mateusieno05@gmail.com, jcrcamara@hotmail.com, ronei@de.ufpb.br

Resumo. *Objetivo: comparar CNNs e Transformers na triagem automatizada de retinopatia diabética (RD) com o BRSET. Método: tarefa binária em 16.266 imagens; pipeline padronizado, aumento de dados, validação cruzada estratificada (5-fold) e teste independente; métricas: acurácia, precisão, sensibilidade, F1 e Kappa. Resultados: o ConvNeXtV2 obteve melhor equilíbrio entre sensibilidade e precisão (Accuracy=0,981; F1=0,848; Kappa=0,838), superando EfficientNetV2M e MaxViT; o SwinV2 apresentou o pior desempenho. Conclusão: O método ConvNeXtV2 mostrou desempenho mais consistente, sugerindo que a escolha deve considerar a natureza das lesões e a representação espacial para maximizar a sensibilidade.*

Abstract. *Objective: To compare CNNs and Transformers for automated diabetic retinopathy (DR) screening using BRSET. Methods: Binary task on 16,266 fundus images; standardized pipeline; data augmentation, stratified 5-fold cross-validation, and an independent test set; metrics: accuracy, precision, recall, F1, and Cohen's Kappa. Results: ConvNeXtV2 achieved the best balance between sensitivity and precision (Accuracy=0.981; F1=0.848; Kappa=0.838), outperforming EfficientNetV2M and MaxViT; SwinV2 had the weakest performance. Conclusion: The method ConvNeXtV2 provided more consistent performance, indicating that choice should account for lesion characteristics and spatial representation to maximize sensitivity.*

1. Introdução

A retinopatia diabética é uma das principais causas de perda visual evitável em adultos em idade produtiva, decorrente de alterações microvasculares associadas ao diabetes mellitus. A doença manifesta-se por lesões retinianas como microaneurismas, hemorragias, exsudatos e neovascularização, podendo evoluir para comprometimento visual irreversível quando o diagnóstico ocorre tardiamente [Ferreira et al., 2024; Santos et al., 2026]. Apesar da existência de protocolos clínicos de rastreio, a crescente demanda por exames oftalmológicos e a escassez de especialistas impõem desafios relevantes aos sistemas de saúde, incluindo o elevado volume de exames e a variabilidade na interpretação das imagens [Akhtar et al., 2025; Teoh et al., 2023].

Nesse contexto, métodos de aprendizagem profunda têm sido investigados para apoiar a triagem automatizada de imagens retinianas. Redes neurais convolucionais demonstraram elevada capacidade de modelar padrões locais e estruturas espaciais, enquanto arquiteturas baseadas em mecanismos de autoatenção passaram a explorar dependências contextuais mais amplas por meio de Vision Transformers [Han et al., 2023]. Estudos recentes indicam o potencial dessas abordagens em tarefas oftalmológicas, embora os resultados variem conforme o desenho experimental, o conjunto de dados e as estratégias de treinamento [Fan et al., 2023; Akhtar et al., 2025].

O Brazilian Retinal Dataset (BRSET) reúne imagens de fundo de olho obtidas em contexto clínico brasileiro, acompanhadas de anotações clínicas e de metadados demográficos [Nakayama et al., 2023]. Neste estudo, a tarefa foi definida como classificação binária em nível de imagem para predizer a presença de retinopatia diabética. O desempenho das arquiteturas é avaliado por meio de acurácia, precisão, sensibilidade, F1-score e coeficiente Kappa [Leite; De Moraes; Lopes, 2022]. Diante desse cenário, este trabalho realiza uma comparação sistemática entre arquiteturas convolucionais e modelos baseados em Transformers sob um protocolo experimental padronizado. Ao controlar a preparação dos dados, o balanceamento, a otimização e a avaliação, o estudo busca analisar como as diferenças arquiteturais influenciam o desempenho discriminativo na triagem automatizada de retinopatia diabética.

2. Materiais e Métodos

O processo metodológico foi organizado em três fases principais: (i) preparação e organização dos dados, (ii) treinamento e otimização das arquiteturas e (iii) avaliação de desempenho. A Figura 1 apresenta o fluxo completo das etapas adotadas no estudo.

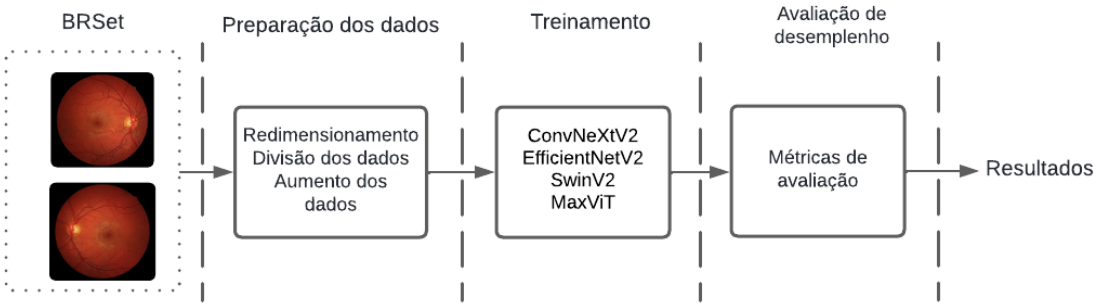

Figura 1. Fluxo de classificação.

2.1. Fase I – Preparação e Organização dos Dados

2.1.1. Conjunto de Dados

O estudo utilizou o Brazilian Multi-Label Ophthalmological Dataset (BRSET), disponibilizado na plataforma PhysioNet, composto por 16.266 imagens coloridas de fundo de olho provenientes de 8.524 pacientes brasileiros [Luis Filipe Nakayama et al., 2023]. As imagens apresentam resolução original aproximada de 951 × 874 pixels e são centradas na região macular, permitindo visualização dos principais arcos vasculares retinianos, conforme ilustrado na Figura 2.

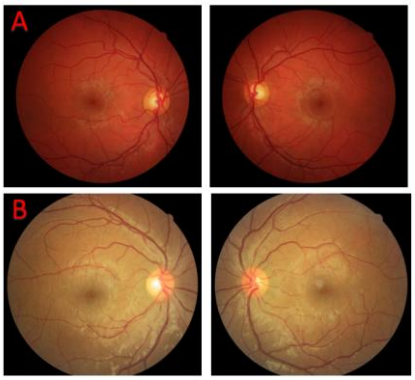

Figura 2. Exemplos de imagens do BRSET: (A) ausência de retinopatia diabética e (B) presença de retinopatia diabética.

Foi formulado um problema supervisionado de classificação binária em nível de imagem, no qual o modelo prediz a presença ou ausência de retinopatia diabética a partir de fotografias retinianas. As amostras foram organizadas em duas classes mutuamente exclusivas: classe 0 (ausência da condição) e classe 1 (presença da condição). A distribuição entre as classes caracteriza um cenário desbalanceado, com maior proporção de exames sem evidência da doença. O conjunto de teste permaneceu completamente independente ao longo de todo o estudo, sendo utilizado exclusivamente na avaliação final. Para treinamento e validação, adotou-se validação cruzada estratificada com cinco folds, preservando a proporção entre as classes em cada partição e reduzindo a variabilidade associada a divisões únicas dos dados. O BRSET constitui um recurso relevante para pesquisas em oftalmologia computacional por reunir imagens obtidas em contexto clínico real e metadados demográficos associados. A Tabela 1 apresenta uma

comparação entre o BRSET e outras bases públicas utilizadas em estudos de detecção de doenças oculares.

Tabela 1. Lista de bases de dados disponíveis de estudos de detecção de doenças oculares.

Dataset	Images	Resolutions	Source
BRSET	16.266	951×874	(Nakayama et al., 2023)
ACRIMA	705	2048×1536	(Elmoufidi; Jai-andaloussi, 2021)
REFUGE	1200	2124×2056, 1634×1634	(Alayón et al., 2023)
RIM-ONE	485	2144×1424	(Fumero Batista et al., 2020)
Drishti-GS1	101	2896×1944	(Sivaswamy et al., 2014)
ORIGA-light	650	3072×2048	(Zhuo Zhang et al., 2010)
sjchoi86-HRF	401	2592×1728	(Camara et al., 2022)
G1020	1.020	1956×1934	(Bajwa et al., 2020)

2.1.2. Preparação e Enriquecimento de Dados

A preparação dos dados e as estratégias de aumento são etapas essenciais para garantir a consistência experimental e a capacidade de generalização dos modelos. Inicialmente, todas as imagens foram organizadas preservando apenas o conteúdo retiniano relevante para a tarefa de classificação. Para assegurar comparabilidade entre as arquiteturas, as imagens foram redimensionadas para 224 × 224 pixels, independentemente do modelo utilizado. Em seguida, aplicou-se a normalização com base na média e no desvio padrão do ImageNet, garantindo compatibilidade com os pesos pré-treinados. Considerando a variabilidade das imagens médicas e o desbalanceamento entre as classes, foram aplicadas estratégias de *data augmentation* exclusivamente no conjunto de treinamento, implementadas dinamicamente a cada *batch*. As transformações incluíram inversão horizontal, rotações controladas, ajustes de brilho e contraste, variações de saturação e matiz, ajuste de gama, desfoque gaussiano e Coarse Dropout.

Para mitigar o desbalanceamento, utilizou-se amostragem ponderada por classe, atribuindo pesos inversamente proporcionais à frequência de cada rótulo. Adicionalmente, foram empregadas as técnicas Mixup e CutMix, aplicadas com probabilidade de 66% por *batch*. Essas estratégias combinam pares de imagens e regiões espaciais durante o treinamento, ampliando a diversidade das amostras e contribuindo para maior estabilidade do modelo sob desequilíbrio de classes.

2.2. Fase II – Treinamento e Otimização dos Modelos

Arquiteturas

Foram avaliadas quatro arquiteturas representativas de diferentes paradigmas estruturais: duas baseadas em redes neurais convolucionais (CNN) — ConvNeXtV2 [Woo et al., 2023] e EfficientNetV2 [Tan e Le, 2021], uma baseada em Transformers, SwinV2 [Liu et al., 2022] e uma arquitetura híbrida que combina convolução e autoatenção, MaxViT [Tu et al., 2022]. Todas foram inicializadas com pesos pré-treinados em ImageNet, conforme disponibilizados nas implementações públicas das respectivas arquiteturas. ConvNeXtV2, uma CNN moderna, incorpora ajustes estruturais que atualizam o desenho das redes convolucionais tradicionais, enquanto EfficientNetV2, também de natureza convolucional, emprega escalonamento estruturado de profundidade, largura e resolução, uma estratégia amplamente utilizada em tarefas de classificação médica. SwinV2, arquitetura baseada em Transformer, e MaxViT, arquitetura híbrida CNN–Transformer, integram mecanismos de atenção capazes de modelar dependências espaciais de maior alcance. SwinV2 utiliza atenção hierárquica com janelas deslocadas, permitindo modelagem contextual com complexidade computacional controlada. MaxViT combina convolução local e atenção global em múltiplas escalas, integrando informações espaciais de curto e de longo alcance. A camada final de cada modelo foi adaptada para gerar duas saídas compatíveis com o problema binário. Essa composição arquitetural permite analisar como diferentes estratégias de extração e integração de características, convolução pura, autoatenção pura e combinação híbrida, influenciam o desempenho discriminativo sob condições experimentais equivalentes.

Estratégia de Treinamento

O treinamento foi conduzido por meio de validação cruzada estratificada com cinco *folds*. Em cada iteração, quatro partições foram destinadas ao treinamento e uma à validação. A otimização foi realizada com o algoritmo AdamW [Loshchilov; Hutter, 2019; Howard et al., 2017], utilizando uma taxa de aprendizagem inicial de 1×10^{-4}. Os hiperparâmetros *weight decay*, *beta1* e *beta2* foram ajustados por meio de uma busca bayesiana com o Optuna [Akiba et al., 2019], visando maximizar a acurácia média na validação. A taxa de aprendizagem foi ajustada dinamicamente pelo *scheduler*, que monitorava a acurácia na validação. A função de perda adotada foi a Focal Loss, parametrizada por $\alpha = 0{,}75$ e $\gamma = 2{,}0$, o que favorece uma maior penalização de exemplos mal classificados. O treinamento foi realizado por até 10 épocas em cada *fold*, com aplicação de Early Stopping baseada na acurácia de validação e paciência de 5 épocas. O modelo com melhor desempenho em cada partição foi armazenado para avaliação posterior [Han et al., 202].

2.3. Fase III – Avaliação de Desempenho

A avaliação final foi realizada no conjunto de teste independente, utilizando o modelo selecionado na etapa de otimização. O desempenho foi mensurado por meio das métricas de acurácia, precisão, sensibilidade, F1-score e coeficiente kappa de Cohen. A

manutenção do conjunto de teste isolado ao longo de todo o processo assegura uma estimativa de generalização em dados não observados durante o treinamento, aproximando a análise de condições clínicas reais [Ferreira et al., 2024; Leite; De Moraes; Lopes, 2022].

3. Resultados

A avaliação do desempenho foi conduzida com base nas métricas previamente definidas, considerando o cenário de desbalanceamento entre as classes. A análise prioriza não apenas a acurácia global, mas também a relação entre sensibilidade, precisão, F1-score e o coeficiente Kappa, indicadores relevantes para a avaliação de modelos aplicados à triagem clínica. A Tabela 2 apresenta os resultados obtidos pelas arquiteturas ConvNeXtV2, MaxViT, EfficientNetV2-M e SwinV2 no conjunto de teste.

Tabela 2. Resultados obtidos pelas arquiteturas ConvNeXtV2, MaxViT, EfficientNetV2-M e SwinV2 no conjunto de teste.

Modelo	Accuracy	Precision	Recall	F1-score	Kappa
ConvNeXtV2 (CNN)	**0,981**	**0,905**	**0,798**	**0,848**	**0,838**
MaxViT (Híbrido)	0,979	0,903	0,774	0,833	0,822
EfficientNetV2-M (CNN)	0,979	0,920	0,750	0,826	0,815
SwinV2 (Transformer)	0,952	0,634	0,690	0,661	0,635

Todos os modelos alcançaram acurácia superior a 0,95. Entretanto, essa métrica isoladamente não descreve adequadamente o desempenho em cenários com desbalanceamento de classes, pois pode ser influenciada pela predominância da classe majoritária [Japkowicz; Stephen, 2002]. Assim, a interpretação deve considerar métricas derivadas da matriz de confusão, como precisão, sensibilidade, F1-score e o coeficiente Kappa [Saito; Rehmsmeier, 2015; Viera; Garrett, 2005]. Com base nesses indicadores, o ConvNeXtV2 apresentou o melhor equilíbrio entre a recuperação da classe positiva e o controle de falsos positivos, com F1-score de 0,848 e Kappa de 0,838. O EfficientNetV2-M, embora tenha apresentado a maior precisão de 0,920, exibiu sensibilidade inferior de 0,750, indicando maior seletividade e maior probabilidade de falsos negativos. O MaxViT apresentou desempenho intermediário, com distribuição relativamente estável entre as métricas. Por sua vez, o SwinV2 apresentou redução mais pronunciada na precisão, na sensibilidade, no F1-score e no Kappa, evidenciando menor desempenho sob desbalanceamento de classes.

4. Discussão

Os resultados indicam que a acurácia isolada não descreve adequadamente o desempenho em cenários de desbalanceamento de classes. A análise conjunta de sensibilidade, precisão, F1-score e coeficiente Kappa revela diferenças consistentes entre as arquiteturas avaliadas. Essas diferenças não se limitam a variações numéricas, mas também refletem

modos distintos de representar e integrar padrões patológicos de pequena escala. A Figura 3 acrescenta evidência qualitativa ao ilustrar como essas diferenças se manifestam no processo de decisão do modelo. Na linha A, composta por verdadeiros negativos, observa-se consistência nas probabilidades atribuídas à classe 0, com níveis elevados de confiança e baixa variação entre as amostras. As imagens apresentam morfologia retiniana preservada, sem sinais microvasculares compatíveis com retinopatia diabética. Esse comportamento indica que o modelo consegue representar adequadamente a estrutura retiniana normal.

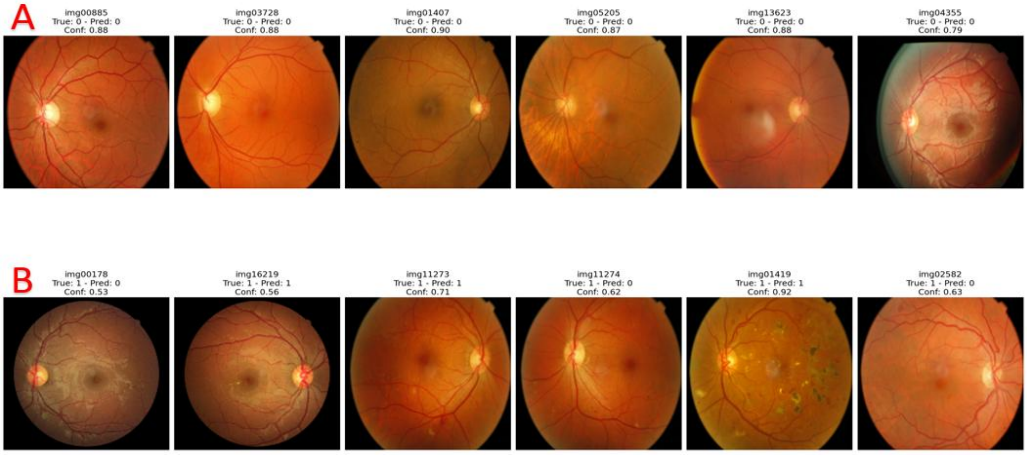

Figura 3. Exemplos de predições do modelo no conjunto de teste do BRSET.

Na linha B, referente à classe positiva, o comportamento difere. Embora imagens com lesões extensas apresentem elevada confiança preditiva, os falsos negativos concentram-se em casos com alterações discretas e altamente localizadas. Microaneurismas isolados e pequenas hemorragias ocupam apenas uma fração limitada da imagem e apresentam contraste limitado em relação ao tecido adjacente. Esse padrão exige a preservação consistente de informação espacial de alta frequência ao longo da hierarquia de representação da rede. A redução progressiva da resolução espacial decorrente de operações de *downsampling*, realizadas por *pooling* ou convoluções com *stride*, pode reduzir a intensidade das ativações associadas a essas estruturas quando elas ocupam regiões muito pequenas da imagem. Nesses casos, a evidência patológica altera pouco as estatísticas globais da imagem, o que dificulta a identificação de padrões discriminativos suficientemente fortes para influenciar a decisão final do modelo.

Esse comportamento ajuda a interpretar as diferenças observadas entre arquiteturas convolucionais e modelos baseados em mecanismos de atenção. Redes convolucionais mantêm uma forte indução local, permitindo a extração de texturas e padrões microvasculares em escalas reduzidas. Modelos baseados predominantemente em atenção dependem de integrações contextuais mais amplas e podem requerer padrões distribuídos para estabilizar a decisão quando treinados com resolução limitada. As probabilidades intermediárias observadas nos falsos negativos da Figura 3 indicam que parte da informação discriminativa foi capturada, embora com intensidade insuficiente para alterar a classe prevista. Estudos recentes mostram que o desempenho de Vision

Transformers em retinografia depende do regime de pré-treino e da forma como a informação de alta resolução é explorada. Estratégias autossupervisionadas, como *masked autoencoders*, e arquiteturas que preservam detalhes estruturais podem melhorar a representação de lesões pequenas [Yang et al., 2024]. Outros trabalhos relatam resultados variáveis na detecção de retinopatia diabética referível, dependendo do protocolo experimental e das métricas adotadas [Wu et al., 2023]. Alguns estudos relatam desempenho elevado de modelos baseados em atenção, enquanto outros mostram que arquiteturas convolucionais e híbridas mantêm desempenho competitivo e interpretação clínica consistente. Os resultados observados no BRSET seguem esse padrão. A arquitetura convolucional ConvNeXtV2 apresentou melhor equilíbrio entre recuperação da classe positiva e controle de erros. A análise qualitativa sugere que os erros se concentram nos estágios iniciais da doença, quando a carga lesional ainda é limitada e as alterações microvasculares ocupam pequenas regiões da imagem. Em protocolos de rastreio populacional, a sensibilidade desempenha papel central, pois o objetivo principal é identificar o maior número possível de casos positivos para encaminhamento diagnóstico [Lestari et al., 2025]. Nesse contexto, o desempenho do ConvNeXtV2 indica maior adequação ao cenário analisado. A menor sensibilidade observada no SwinV2 sugere maior risco de não identificação de casos com alterações discretas.

A distribuição das probabilidades indica que o limiar de decisão constitui um elemento relevante na aplicação clínica desses modelos. Em ambientes de triagem, o limiar pode ser ajustado para priorizar sensibilidade, de acordo com a estratégia assistencial adotada. Esse aspecto mostra que o desempenho preditivo não depende apenas da arquitetura, mas também da forma como as probabilidades geradas pelo modelo são integradas ao fluxo clínico. De forma geral, os resultados indicam que diferenças arquiteturais influenciam a capacidade de preservar e integrar evidências microvasculares altamente localizadas. A comparação conduzida sob um protocolo experimental uniforme mostra que o comportamento discriminativo está associado à forma como cada arquitetura representa a informação espacial em múltiplas escalas. Assim, a escolha arquitetural para sistemas de triagem automatizada deve considerar não apenas métricas globais de desempenho, mas também a natureza morfológica das lesões, a resolução das imagens e as estratégias de treinamento adotadas.

5. Considerações Finais

Este estudo comparou arquiteturas convolucionais modernas e Vision Transformers para a triagem automatizada de retinopatia diabética no BRSET, utilizando um protocolo experimental padronizado, com validação cruzada estratificada e teste independente. Em cenário de desbalanceamento de classes, o ConvNeXtV2 apresentou o melhor equilíbrio entre sensibilidade e precisão, com valores superiores de F1-score e coeficiente Kappa, enquanto o SwinV2 apresentou desempenho inferior. Esses resultados indicam que diferenças arquiteturais influenciam a capacidade de detectar lesões retinianas em tarefas de triagem clínica e reforçam a necessidade de avaliar modelos com métricas além da acurácia. Do ponto de vista aplicado, os resultados indicam que o ConvNeXtV2 constitui

uma alternativa adequada para sistemas automatizados de triagem de retinopatia diabética. Além disso, o estudo demonstra que decisões metodológicas relacionadas ao pré-processamento, aumento de dados, particionamento estratificado e uso de conjunto de teste independente influenciam diretamente o desempenho observado e devem integrar protocolos de avaliação em visão computacional aplicada à oftalmologia.

Referências

Akhtar, S. et al. (2025) A deep learning based model for diabetic retinopathy grading. Scientific Reports, v. 15, n. 1, p. 3763.

Akiba, T. et al. (2019) Optuna: A Next-generation Hyperparameter Optimization Framework. arXiv.

Alayón, S. et al. (2023) Comparison of the Performance of Convolutional Neural Networks and Vision Transformer-Based Systems for Automated Glaucoma Detection with Eye Fundus Images. Applied Sciences, v. 13, n. 23, p. 12722.

Bajwa, M. N. et al. (2020) G1020: A Benchmark Retinal Fundus Image Dataset for Computer-Aided Glaucoma Detection. In: 2020 International Joint Conference on Neural Networks (IJCNN). IEEE.

Camara, J. et al. (2022) Literature Review on Artificial Intelligence Methods for Glaucoma Screening, Segmentation, and Classification. Journal of Imaging, v. 8, n. 2, p. 19.

Cohen, J. (1960) A Coefficient of Agreement for Nominal Scales. Educational and Psychological Measurement, v. 20, n. 1, p. 37–46.

D S, R.; Saji, K. S. (2025) Hybrid deep learning framework for diabetic retinopathy classification with optimized attention AlexNet. Computers in Biology and Medicine, v. 190, p. 110054.

Elmoufidi, A.; Jai-andaloussi, S. (2021) CNN with Multiple Input for automatic glaucoma assessment using Fundus Images. In Review.

Fan, R. et al. (2023) Detecting Glaucoma from Fundus Photographs Using Deep Learning without Convolutions. Ophthalmology Science, v. 3, n. 1, p. 100233.

Ferreira, J. S. et al. (2024) Application of Vision Transformers in the Early Detection of Excavation in the BRSET Base. In: Proceedings of DSAI 2024. ACM.

Fumero Batista, F. J. et al. (2020) RIM-ONE DL: A Unified Retinal Image Database for Assessing Glaucoma Using Deep Learning. Image Analysis & Stereology, v. 39, n. 3, p. 161–167.

Han, K. et al. (2023) A Survey on Vision Transformer. IEEE Transactions on Pattern Analysis and Machine Intelligence, v. 45, n. 1, p. 87–110.

Howard, A. G. et al. (2017) MobileNets: Efficient Convolutional Neural Networks for Mobile Vision Applications. arXiv.

Japkowicz, N.; Stephen, S. (2002) The class imbalance problem: A systematic study. Intelligent Data Analysis, v. 6, n. 5, p. 429–449.

Leite, D. R. A.; De Moraes, R. M.; Lopes, L. W. (2022) Different Performances of Machine Learning Models to Classify Dysphonic and Non-Dysphonic Voices. Journal of Voice.

Lestari, Y. D. et al. (2025) Diabetic retinopathy screening model in low and middle-income countries: a scoping review. BMC Public Health, v. 25, n. 1, p. 4210.

Liu, Z. et al. (2022) Swin Transformer V2: Scaling Up Capacity and Resolution. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), p. 12009–12019.

Loshchilov, I.; Hutter, F. (2019) Decoupled Weight Decay Regularization. ICLR 2019.

Nakayama, L. F. et al. (2023) BRSET: A Brazilian Multilabel Ophthalmological Dataset of Retina Fundus Photos. Scientific Data, v. 10, Article 283. https://doi.org/10.1038/s41597-023-02124-4

Saito, T.; Rehmsmeier, M. (2015) The Precision-Recall Plot Is More Informative than the ROC Plot When Evaluating Binary Classifiers on Imbalanced Datasets. PLOS ONE, v. 10, n. 3, p. e0118432.

Santos, C. et al. (2026) Brazilian Dataset for Retinal Lesion Analysis: A Deep Learning Diagnostic Pipeline. Journal of Health Informatics, v. 18.

Sivaswamy, J. et al. (2014) Drishti-GS: Retinal Image Dataset for Optic Nerve Head (ONH) Segmentation. In: Proceedings of the 2014 IEEE 11th International Symposium on Biomedical Imaging (ISBI), p. 53–56.

Tan, M.; Le, Q. V. (2021) EfficientNetV2: Smaller Models and Faster Training. In: Proceedings of the 38th International Conference on Machine Learning (ICML), p. 10096–10106.

Teoh, C. S. et al. (2023) Variability in Grading Diabetic Retinopathy Using Retinal Photography and Its Comparison with an Automated Deep Learning Diabetic Retinopathy Screening Software. Healthcare, v. 11, n. 12, p. 1697.

Tu, Z. et al. (2022) MaxViT: Multi-axis Vision Transformer. In: Computer Vision – ECCV 2022. Lecture Notes in Computer Science. Springer, v. 13684, p. 459–479.

Viera, A. J.; Garrett, J. M. (2005) Understanding Interobserver Agreement: The Kappa Statistic. Family Medicine, v. 37, p. 360–363.

Woo, S. et al. (2023) ConvNeXt V2: Co-designing and Scaling ConvNets with Masked Autoencoders. In: 2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). IEEE.

Wu, J.-H. et al. (2023) Vision transformers: The next frontier for deep learning-based ophthalmic image analysis. Saudi Journal of Ophthalmology, v. 37, n. 3, p. 173–178.

Yang, Y. et al. (2024) Vision transformer with masked autoencoders for referable diabetic retinopathy classification based on large-size retina image. PLOS ONE, v. 19, n. 3, p. e0299265.

Zhang, Z. et al. (2010) ORIGA-light: An Online Retinal Fundus Image Database for Glaucoma Analysis and Research. In: Proceedings of the 2010 Annual International Conference of the IEEE Engineering in Medicine and Biology Society (EMBC), p. 3065–3068.