Desenvolvimento de um Protótipo para Triagem Automatizada de Glaucoma por Deep Learning Integrado ao Whatsapp

Gabriele R. Wanzeler[1], Otávio N. Teixeira[1]
[1]Faculdade de Engenharia de Computação - Universidade Federal do Pará
Campus Universitário de Tucuruí - PA - Brasil

gabrielewanzeler000@gmail.com, otaviont@ufpa.br

Abstract. *This work evaluates a ResNet50 model with transfer learning for automated glaucoma detection in fundoscopy images, integrated into a WhatsApp prototype for healthcare professionals. The model was trained on the ACRIMA and RIM-ONE datasets. On ACRIMA, it achieved 98.28% sensitivity, F1-score of 0.982, and AUC of 0.992; on RIM-ONE, 83.33% sensitivity, F1-score of 0.869, and AUC of 0.949, reflecting class imbalance. The prototype achieved accuracy rates of 83.33% and 80%, with average confidence of 94.80% and 87.31%, respectively. Results show that data balance directly influences performance, reinforcing the potential of deep learning as a support tool for glaucoma screening.*

Resumo. *Este trabalho avalia um modelo ResNet50 com transfer learning para detecção automatizada de glaucoma em imagens de fundoscopia, integrado a um protótipo via WhatsApp para profissionais de saúde. O modelo foi treinado nos datasets ACRIMA e RIM-ONE. No ACRIMA, alcançou sensibilidade de 98,28%, F1-score de 0,982 e AUC de 0,992; no RIM-ONE, sensibilidade de 83,33%, F1-score de 0,869 e AUC de 0,949, reflexo do desbalanceamento de classes. O protótipo obteve taxa de acerto de 83,33% e 80%, com confiança média de 94,80% e 87,31%, respectivamente. Os resultados evidenciam que o balanceamento dos dados influencia diretamente o desempenho, reforçando o potencial do aprendizado profundo como ferramenta de apoio à triagem de glaucoma.*

1.Introdução

A busca pela automatização de processos voltados à resolução de problemas clínicos tem avançado de forma significativa com o desenvolvimento de técnicas baseadas em *Deep Learning*, as quais possibilitam a análise automática de grandes volumes de dados e a extração de padrões relevantes sem a necessidade de intervenção manual especializada [LeCun 2015].

Apesar dos avanços tecnológicos, a interpretação e análise de imagens do fundo do olho ainda é um método exaustivo e suscetível a erros [Tamim et al. 2021]. Dessa forma, o uso dessas técnicas de automatização tem impactado de maneira direta o desenvolvimento de sistemas de triagem e de apoio à decisão em saúde, especialmente em cenários caracterizados por grande volume de exames e limitação de recursos humanos especializados. Soluções baseadas em aprendizado de máquina possibilitam a análise preliminar de dados clínicos e de imagens médicas de forma rápida, contribuindo para a priorização de casos suspeitos, particularmente em contextos de telemedicina. Nesse sentido, tais sistemas não têm como objetivo substituir a avaliação médica, mas atuar como ferramentas auxiliares capazes de otimizar fluxos de atendimento, reduzir atrasos diagnósticos e apoiar estratégias de rastreamento em larga

escala. No Brasil, o WhatsApp Messenger® destaca-se pela adoção massiva entre profissionais de saúde e pacientes, tornando-se um canal acessível e familiar [Xavier 2024]. A integração entre modelos de inteligência artificial e o WhatsApp pode, portanto, ampliar o alcance de ferramentas de triagem, reduzindo barreiras geográficas e socioeconômicas.

A triagem automatizada traz vantagens que vão além de tornar o trabalho mais rápido. Ela ajuda a manter as avaliações mais consistentes, diminui os erros que podem acontecer por causa do cansaço ou da sobrecarga dos profissionais, e também contribui para uma utilização mais eficiente dos recursos nos hospitais. Implementações concretas no cenário brasileiro evidenciam a viabilidade dessas soluções, como é o caso da Neomed que criou um sistema com inteligência artificial que pode realizar a triagem de eletrocardiogramas em até 13 segundos, detectando anormalidades cardiovasculares críticas [Neomed 2025]. A enfermagem consolida-se como mediadora indispensável nesse processo, sendo responsável pela validação clínica e humanização do atendimento automatizado [Paiva, Gomes e Takaoka 2025].

Nesse contexto, a aplicação dessas abordagens na detecção automatizada de doenças oculares, como o glaucoma, que se caracteriza por uma neuropatia óptica com repercussão característica no campo visual, é classificado como a principal causa de cegueira irreversível e a segunda maior causa de cegueira global, ficando atrás apenas da catarata [WHO 2019]. Em 2020, estimou-se que a doença afetou globalmente 64 milhões de indivíduos com idades entre 40 e 80 anos, e espera-se que esse número suba para cerca de 111,8 milhões devido ao envelhecimento da população até o ano de 2040 [Tham et al. 2014].

A partir disso, este trabalho tem como foco a avaliação do desempenho de um modelo baseado na arquitetura ResNet50 aplicado à detecção automatizada de glaucoma, bem como o desenvolvimento e a análise de um protótipo funcional integrado ao modelo. Para isso, a rede foi treinada e validada separadamente em dois conjuntos públicos de imagens de fundo de olho, ACRIMA e RIM-ONE, possibilitando a análise individual dos resultados obtidos em cada base. A comparação entre os desempenhos do modelo e do protótipo em cada *dataset* permite investigar a influência das características dos dados tanto na capacidade de classificação quanto no comportamento do sistema desenvolvido.

2. Fundamentação Teórica

2.1. Doenças Oculares

As doenças oculares são condições oftalmológicas que representam um conjunto amplo de alterações que comprometem desde estruturas mais externas como retina e nervo óptico, onde a médio e longo prazo podem causar, entre outras coisas, dificuldade na visão e até mesmo, em casos mais graves, a cegueira [Bragança et al. 2022]. Muitas dessas condições evoluem de maneira lenta, o que faz com que o paciente só perceba alguma mudança quando a perda visual já é significativa. Dados da Organização Mundial da Saúde indicam que uma parcela importante dos casos de deficiência visual poderia ser evitada se o diagnóstico fosse feito mais cedo [WHO 2019].

Em certos casos, fatores como genética, condições sistêmicas, envelhecimento ou certos hábitos podem contribuir para o surgimento das alterações. O glaucoma, a catarata, a retinopatia diabética e a degeneração macular relacionada à idade (DMRI) são algumas das doenças mais comuns, cada uma afetando o olho de maneira diferente e exigindo métodos específicos de avaliação. O uso de exames de imagem, como retinografia e tomografia de coerência óptica, possibilita a detecção de alterações muito precoces nas estruturas oculares, aumentando as chances de intervenção antes que danos mais graves se desenvolvam. Rezende (2020) ressalta que o monitoramento regular e a análise atenta dessas imagens são fundamentais para prevenir que o progresso das doenças seja inadvertido.

2.1.1. Glaucoma

Segundo dados da *World Health Organization* (WHO), o glaucoma é responsável por mais de 4,5 milhões de ocorrências de perda total de visão no mundo [WHO]. O tipo mais comum é o glaucoma de ângulo aberto, que afeta tanto pessoas brancas quanto negras e representa um problema de drenagem interno do olho, fazendo com que a pressão intraocular fique elevada, com consequente lesão do nervo óptico. Já o glaucoma de ângulo fechado é mais frequente entre as populações asiáticas [Baudouin 2021], sendo o segundo tipo mais comum, caracterizado pela obstrução da abertura do sistema de drenagem do olho.

O glaucoma é uma doença ocular causada principalmente pela elevação da pressão intraocular que provoca lesões no nervo óptico, além de gerar danos às células da retina e alterar a forma do disco óptico, tendo como consequência o comprometimento visual. É uma doença que costuma avançar sem provocar dor ou incômodo perceptível, o que explica a quantidade de casos diagnosticados tardiamente. A progressão da doença pode ser observada através de alterações morfológicas no disco óptico, como demonstrado na Figura 1. Em um nervo óptico saudável, a escavação (área central mais clara do disco) apresenta uma relação normal com a borda do disco óptico. À medida que o glaucoma progride, essa escavação aumenta devido à perda progressiva de fibras nervosas, resultando em um afinamento da borda do disco. No glaucoma moderado, já é possível observar um aumento significativo da escavação, enquanto no glaucoma avançado, a escavação torna-se muito pronunciada, indicando perda extensa de tecido neural e comprometimento severo da função visual.

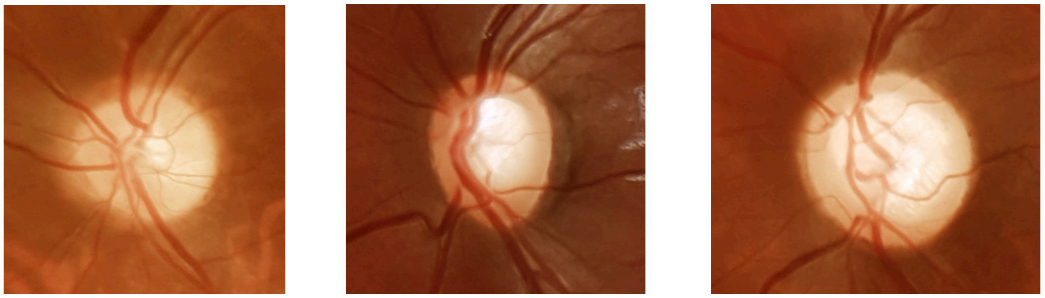

Figura 1. Comparação entre nervo óptico saudável e com glaucoma

O diagnóstico do glaucoma baseia-se no exame da pressão intraocular, na avaliação do nervo óptico e no exame de campo visual. A análise do campo visual, por exemplo, ajuda a identificar padrões de perda de sensibilidade que são comuns na doença. Já os exames estruturais, como a tomografia de coerência óptica do nervo óptico (CT), permitem visualizar mudanças no nervo óptico. Segundo Weinreb (2014), até alterações muito pequenas já podem indicar o começo de danos causados pelo glaucoma, por isso é fundamental fazer exames regularmente.

2.2. Deep Learning

Deep learning (Aprendizado Profundo) permite que modelos computacionais que são compostos de múltiplas camadas de processamento aprendam representações de dados com múltiplos níveis de abstração. Segundo Goodfellow (2016), o termo "profundo" refere-se ao número de camadas através das quais os dados são transformados, permitindo que o modelo aprenda características cada vez mais abstratas e complexas.

Segundo LeCun, Bengio e Hinton (2015), o *Deep Learning* representa um avanço significativo em relação aos métodos tradicionais de aprendizado de máquina, principalmente por sua capacidade de aprender automaticamente representações relevantes a partir dos dados brutos, sem a necessidade de extração manual de características. Essa propriedade é especialmente importante em domínios complexos, como o processamento de imagens médicas, nos quais a definição explícita de atributos discriminativos pode ser limitada ou imprecisa.

No campo da visão computacional, redes profundas são capazes de modelar padrões visuais em diferentes níveis de abstração, aprendendo inicialmente características de baixo nível, como bordas e variações de intensidade, e progressivamente representações mais complexas relacionadas à forma e à estrutura dos objetos presentes na imagem. Esse comportamento hierárquico torna o *Deep Learning* particularmente eficaz para aplicações em saúde, onde pequenas variações visuais podem indicar alterações patológicas relevantes [Litjens et al. 2017].

Aplicações recentes demonstram que modelos baseados em *Deep Learning* têm alcançado desempenho elevado em tarefas de classificação de imagens oftalmológicas, incluindo a detecção automatizada de glaucoma a partir de imagens de fundo de olho. De acordo com Ting et al. (2017), esses modelos apresentam potencial para atuar como ferramentas de apoio à decisão clínica, especialmente em cenários de triagem, nos quais há grande volume de exames e necessidade de identificação precoce da doença.

2.2.1. Redes Neurais Convolucionais (CNNs)

Goodfellow (2016) descrevem as Redes Neurais Convolucionais como estruturas semelhantes às Redes Neurais Artificiais, mas que possuem um número maior de camadas e operações. Em uma Rede Convolucional, cada camada tem a função de captar certos detalhes dos dados de entrada. O fluxo de informações acontece de uma camada para a próxima, com a saída de uma camada servindo como entrada para a próxima etapa da rede. Ao contrário das redes neurais totalmente conectadas, CNNs por sua vez aproveitam a estrutura espacial dos dados por meio de três características essenciais: conectividade local, compartilhamento de pesos e invariância translacional.

Uma abordagem frequentemente empregada para otimizar o treinamento de CNNs é o *transfer learning* (aprendizado por transferência), técnica que consiste em utilizar modelos pré-treinados em grandes bases de dados e adaptá-los para tarefas específicas, reduzindo o tempo de treinamento e melhorando o desempenho. Algoritmos utilizando aprendizado por transferência também foram avaliados em diversos outros trabalhos, como no estudo de Shibata et al. (2018), nos quais conjuntos de dados privados foram empregados e a acurácia obtida superou 90%. Christopher et al. (2018) investigou três arquiteturas distintas de aprendizado profundo. Para cada arquitetura, duas versões diferentes foram avaliadas: aprendizado nativo e aprendizado por transferência. Em todos os casos, os autores demonstraram que o aprendizado por transferência pode aprimorar o desempenho e reduzir o tempo de treinamento dos algoritmos.

Apesar de seu elevado poder de representação, redes neurais convolucionais profundas estão sujeitas ao problema de *overfitting* (sobreajuste), especialmente quando treinadas com conjuntos de dados reduzidos. Nesses casos, o modelo tende a memorizar padrões específicos do conjunto de treinamento, comprometendo sua capacidade de generalização para novos dados. Segundo Goodfellow (2016), estratégias como o uso de aprendizado por transferência, regularização e aumento artificial de dados são amplamente adotadas para minimizar esse efeito, tornando o treinamento mais estável e confiável.

2.2.2. ResNet50

No trabalho de He (2016) foi introduzida a arquitetura *Residual Network* (ResNet) com o objetivo de resolver limitações observadas no treinamento de redes neurais muito profundas, especialmente o problema do desvanecimento do gradiente. Em arquiteturas convencionais, o aumento do número de camadas nem sempre resulta em melhor desempenho, podendo inclusive degradar a capacidade de aprendizado do modelo. A ResNet contorna esse problema por meio do uso de conexões residuais, também conhecidas como skip connections, que permitem a propagação direta da informação entre camadas não consecutivas.

A ResNet50 é uma variação dessa arquitetura composta por cinquenta camadas profundas organizadas em blocos residuais. Cada bloco é projetado para aprender uma função residual, isto é, a diferença entre a entrada e a saída desejada, em vez de aprender diretamente a transformação completa. Essa estratégia facilita o processo de otimização da rede e possibilita o treinamento eficaz de modelos profundos, mantendo estabilidade mesmo com o aumento da complexidade arquitetural [He 2016].

Conforme apresentado na Figura 2, a arquitetura se organiza em cinco estágios sequenciais que processam progressivamente as características da imagem de entrada. O fluxo de informação inicia-se com operações de pré-processamento e segue por múltiplos blocos residuais, onde as conexões de atalho (*skip connections*) possibilitam que o gradiente seja retropropagado de forma mais eficiente através das camadas profundas. Essa característica arquitetural é fundamental para moderar o problema de degradação do desempenho em redes muito profundas, permitindo que a ResNet50 aprenda representações hierárquicas complexas sem comprometer a convergência do treinamento.

Além disso, a ResNet50 é frequentemente adotada em conjunto com estratégias de aprendizado por transferência, nas quais o modelo é inicializado com pesos pré-treinados em grandes bases de dados, como o ImageNet, e posteriormente ajustado para domínios específicos, abordagem particularmente vantajosa em aplicações médicas com conjuntos de dados rotulados limitados.

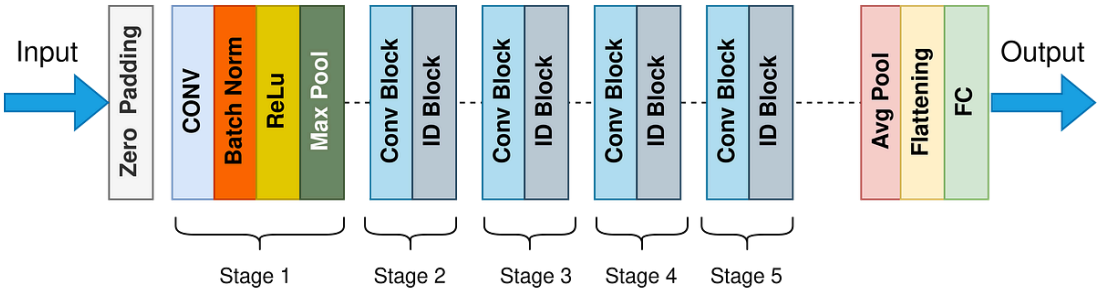

Figura 2. Arquitetura ResNet50

2.2.3. Métricas de Avaliação

No contexto de aprendizado de máquina e redes neurais, a avaliação do desempenho de modelos de classificação é uma etapa fundamental para verificar sua capacidade de generalização e confiabilidade.

As métricas detalhadas a seguir são amplamente empregadas na literatura de *deep learning* e redes neurais convolucionais (CNNs) para a avaliação de modelos de classificação e foram selecionadas neste trabalho por fornecerem uma análise complementar e abrangente do desempenho do modelo.

- **Matriz de Confusão**: Segundo Fawcett (2006), a base para essa análise é a Matriz de Confusão, que permite visualizar o desempenho do algoritmo em quatro categorias fundamentais: Verdadeiro Positivo (VP), Falso Positivo (FP), Verdadeiro Negativo (VN) e Falso Negativo (FN). Ela ilustra o desempenho de um modelo ao comparar as classes previstas com as classes reais, permitindo observar diretamente erros de classificação (FP e FN) que impactam diretamente na eficácia do modelo em aplicações críticas de saúde. Essa representação é amplamente utilizada em tarefas de classificação supervisionada, servindo como base para o cálculo das principais métricas de avaliação adotadas em redes neurais.
- **Acurácia:** Indica a performance geral do modelo, sendo a razão entre as previsões corretas e o total de amostras avaliadas. No entanto, conforme destacado por Goodfellow (2016), a acurácia pode ser uma métrica enganosa em conjuntos de dados desbalanceados, situação comum em problemas reais de classificação, pois não reflete adequadamente o desempenho do modelo sobre a classe minoritária.

$$Acurácia = \frac{VP + VN}{VP + VN + FP + FN}$$

- **Sensibilidade (Recall):** A sensibilidade, também denominada recall ou taxa de verdadeiros positivos (True Positive Rate), mede a capacidade do modelo em identificar corretamente as instâncias pertencentes à classe positiva. Neste trabalho, considera-se como classe positiva a presença de glaucoma. Matematicamente, a sensibilidade é definida como:

$$Sensibilidade = \frac{VP}{VP + FN}$$

onde VP representa os verdadeiros positivos e FN os falsos negativos. Segundo Powers (2011), essa métrica é fundamental em cenários nos quais a não detecção da classe positiva representa um erro crítico. No contexto médico, falsos negativos correspondem a casos da doença que não foram identificados pelo modelo, podendo comprometer o rastreamento precoce.

- **Precisão:** A precisão quantifica a proporção de verdadeiros positivos entre todas as instâncias classificadas como positivas pelo modelo. É definida por:

$$Precisão = \frac{VP}{VP + FP}$$

onde FP representa os falsos positivos. Conforme discutido por Powers (2011), a precisão indica o grau de confiabilidade das predições positivas realizadas pelo classificador. Em aplicações clínicas, alta precisão reduz a ocorrência de encaminhamentos indevidos decorrentes de classificações incorretas.

- **Especificidade**: A especificidade, também conhecida como taxa de verdadeiros negativos (True Negative Rate), mede a capacidade do modelo em identificar corretamente as instâncias negativas. É definida como:

$$Especificidade = \frac{VN}{VN + FP}$$

onde VN representa os verdadeiros negativos. Na área da saúde, a especificidade é amplamente empregada na avaliação de testes diagnósticos, pois indica a capacidade do sistema em reconhecer corretamente indivíduos saudáveis, evitando classificações equivocadas. Essa métrica é especialmente relevante quando o impacto de falsos positivos deve ser controlado.

- **Curva ROC:** A curva ROC (*Receiver Operating Characteristic*) é um gráfico usado para mostrar a capacidade de diagnóstico de classificadores binários. A curva ROC mostra a relação entre sensibilidade e especificidade. Classificadores que geram curvas mais próximas do canto superior esquerdo indicam um melhor desempenho. Como referência, espera-se que um classificador aleatório gere pontos ao longo da diagonal.

- **F1 - Score:** O F1-score é uma métrica que combina precisão e sensibilidade em uma única medida, sendo especialmente útil em problemas de classificação com possível desbalanceamento entre classes. Ele corresponde à média harmônica entre precisão e sensibilidade, penalizando valores discrepantes entre essas duas métricas. Sua formulação é dada por:

$$F1 = 2 * \frac{Precisão * Sensibilidade}{Precisão + Sensibilidade}$$

Por utilizar a média harmônica, o F1-score assume valores elevados apenas quando tanto a precisão quanto a sensibilidade apresentam desempenho satisfatório. Conforme discutido por Powers (2011), essa métrica é particularmente adequada em cenários nos quais há necessidade de equilibrar a identificação correta da classe positiva e a confiabilidade das predições realizadas pelo modelo.

3. Trabalhos Relacionados

Paiva, Gomes e Takaoka (2025) mostram que a inteligência artificial aumenta a precisão da triagem em até 30%, além de diminuir os tempos de espera em serviços de emergência entre 20% e 40%. Das Graças Mendes Júnior et al. (2025) destaca que os sistemas de inteligência artificial ajudam a diminuir o tempo necessário para fazer as classificações, especialmente quando comparados aos processos feitos manualmente. Além disso, algoritmos como redes neurais têm uma sensibilidade maior na hora de identificar casos mais graves. Pesquisas mostram que modelos baseados em inteligência artificial e aprendizado de máquina ajudam a tornar a priorização do atendimento mais precisa, melhorando a distribuição dos recursos e acelerando o tempo até a intervenção [Braz 2025].

No contexto específico da oftalmologia, múltiplos estudos comprovam a eficácia do *deep learning* na detecção automatizada de patologias oculares. Gulshan et al. (2016) desenvolveu um algoritmo de *deep learning* que alcançou sensibilidade de 97,5% e especificidade de 93,4% na detecção de retinopatia diabética moderada e severa, processando 128.175 imagens de retina validadas por oftalmologistas. Para glaucoma especificamente, Li et al. (2018) criou um sistema baseado em redes neurais convolucionais que atingiu AUC (*Area Under the Curve*) de 0,986 na identificação de neuropatia óptica glaucomatosa, superando métodos tradicionais de triagem. Ting et al. (2017) validou um sistema de *deep learning* em 494.661 imagens de retina de populações multiétnicas, demonstrando sua aplicabilidade em diferentes contextos populacionais com AUC superior a 0,93 para múltiplas doenças oculares. Mars e Scott (2017) sistematizaram 49 estudos sobre o uso do WhatsApp em ambientes clínicos, identificando aplicações bem-sucedidas em teleconsulta, triagem e acompanhamento pós-operatório em 14 países.

4. Metodologia

A presente pesquisa caracteriza-se como quantitativa, de natureza aplicada e com delineamento experimental. De acordo com Prodanov e Freitas (2013), a pesquisa quantitativa utiliza dados mensuráveis e análise estatística para interpretação dos

resultados. Neste trabalho, essa abordagem está relacionada à avaliação do desempenho tanto do modelo treinado quanto do protótipo, por meio de métricas quantitativas, permitindo a comparação dos resultados obtidos para cada *dataset* utilizado, além de tratar também de uma pesquisa aplicada, pois visa ao desenvolvimento de uma solução prática para apoio à triagem de glaucoma.

A Figura 3 apresenta a primeira fase do sistema proposto, ilustrando o fluxo inicial com o treinamento do modelo de *deep learning*. O processo se inicia com a utilização de duas bases de dados públicas de imagens de fundo de olho, ACRIMA e RIM-ONE, para o treinamento de modelos baseados na arquitetura ResNet50.

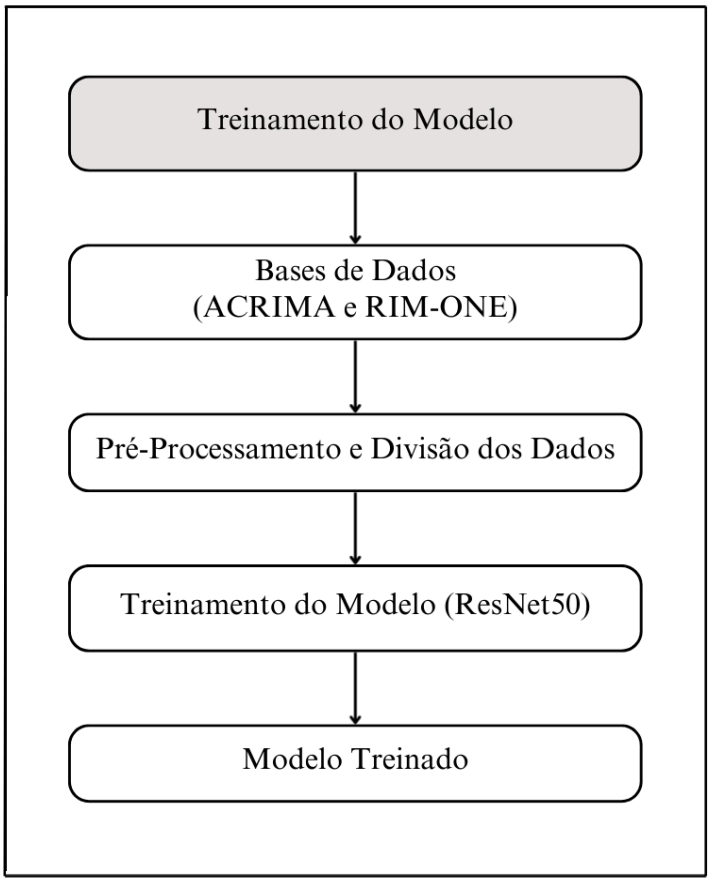

Figura 3. Fluxograma da fase 1 da metodologia proposta

Após o treinamento, conforme a Figura 4, os modelos são salvos e integrados a uma API desenvolvida em Flask, responsável por realizar a comunicação entre o modelo de classificação e a plataforma de mensagens. O usuário final interage com o sistema enviando imagens de fundo de olho através do WhatsApp, recebendo como resposta a classificação automatizada indicando a presença ou ausência de sinais de glaucoma.

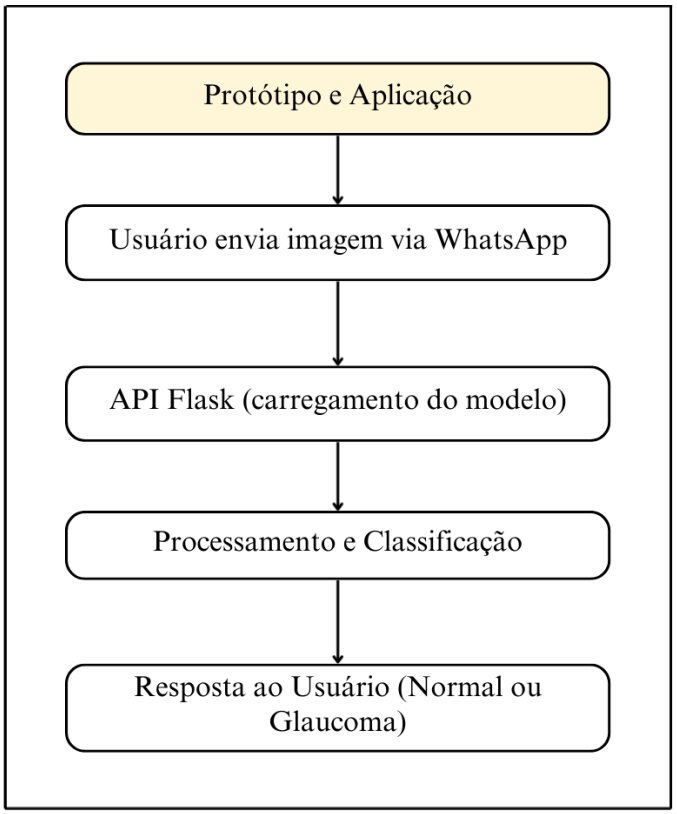

Figura 4. Fluxograma da fase 2 da metodologia proposta

As seções seguintes detalham cada etapa deste processo, incluindo a descrição das bases de dados utilizadas, as técnicas de pré-processamento aplicadas, a arquitetura do modelo de *deep learning*, os procedimentos de treinamento e validação, além da implementação técnica da integração com o WhatsApp.

4.1. Ambiente de experimentos

O desenvolvimento deste trabalho envolveu três ambientes computacionais distintos, cada um escolhido conforme as demandas específicas de cada etapa do protótipo proposto.

Ambiente 1: Treinamento do Modelo

O treinamento do modelo de classificação binária foi realizado na plataforma Kaggle, o notebook utilizado contou com uma unidade da GPU Nvidia Tesla P100, acompanhada de 4 núcleos de CPU e 29 GB de memória RAM, sendo também disponibilizado um espaço em disco de 20 GB. Cada sessão de uso da GPU conta com o tempo máximo de 12 horas, sendo 30 horas de uso o máximo liberado para as GPUs. Neste ambiente, foram armazenados e processados os dois conjuntos de dados utilizados no trabalho (ACRIMA e RIM-ONE) e todo o código de treinamento foi implementado em Python.

As principais bibliotecas usadas nessa etapa incluem TensorFlow e Keras para a construção e treinamento do modelo ResNet50, NumPy utilizada para manipulação de arrays multidimensionais e operações matemáticas necessárias no pré-processamento

das imagens, a biblioteca OpenCV aplicada para operações de processamento de imagens, a biblioteca Scikit-learn empregada na divisão estratificada dos dados em conjuntos de treinamento, validação e teste, além do cálculo de métricas de avaliação e Matplotlib e Seaborn que foram implementadas para a geração de gráficos e representações visuais dos resultados experimentais, incluindo matriz de confusão e curva ROC.

Ambiente 2: API de Classificação

A API (*Application Programming Interface*), de classificação foi desenvolvida em Python utilizando o *framework* Flask e executada no ambiente Google Colaboratory, também chamado de Google Colab. Esse ambiente foi responsável por hospedar o modelo de aprendizado profundo previamente treinado, bem como por processar as requisições de classificação de imagens. Para permitir o acesso externo à API a partir do servidor local, foi utilizada a ferramenta ngrok, que fornece uma URL pública temporária por meio de túneis HTTP seguros.

Ambiente 3: WhatsApp

A integração com o WhatsApp foi implementada como uma aplicação local em Node.js com TypeScript, executada em um notebook pessoal equipado com processador Ryzen 5 5600 G, 16 GB de memória RAM e 512 GB de armazenamento em disco, sem placa de vídeo dedicada. O desenvolvimento foi realizado no editor Visual Studio Code. A integração com o aplicativo de mensagens foi realizada por meio da biblioteca Baileys, responsável pelo gerenciamento da comunicação com a plataforma. Para a comunicação com a API de classificação, foram utilizadas as bibliotecas Axios, para requisições HTTP, e Form-Data para o envio de imagens.

4.2. Arquitetura Utilizada

Para a classificação das imagens de fundo de olho, foi empregada a arquitetura ResNet50 com a técnica de *transfer learning*, aproveitando pesos pré-treinados na base ImageNet. Conforme descrito na seção 2.2.2, a ResNet50 é uma rede neural convolucional profunda que utiliza conexões residuais para facilitar o treinamento de redes com maior número de camadas, sendo amplamente adotada em tarefas de classificação de imagens médicas.

A implementação do transfer learning foi realizada através do congelamento das camadas iniciais da rede base, mantendo apenas as últimas 20 camadas treináveis. Essa estratégia, conhecida como *fine-tuning* (*ajuste fino*), permite que o modelo preserve os recursos de baixo nível aprendidos no ImageNet, como detecção de bordas e texturas, enquanto adapta as camadas mais profundas para as características específicas das imagens de fundo de olho relacionadas ao glaucoma.

Sobre a rede base ResNet50, foi adicionado um cabeçalho de classificação personalizado composto por uma camada de *Global Average Pooling*, seguida de três camadas densas totalmente conectadas com 512, 256 e 128 neurônios, respectivamente. Entre essas camadas, foram inseridas camadas de *Batch Normalization* para estabilizar o treinamento e camadas de *Dropout* com taxas de 0.4, 0.3 e 0.2 para prevenir *overfitting*. A camada de saída utiliza um único neurônio com função de ativação sigmoide, adequada para a tarefa de classificação binária entre imagens normais e com glaucoma.

4.2.1 Configurações de Treinamento

O treinamento do modelo foi conduzido ao longo de 100 épocas, utilizando batches de 16 imagens. Para otimização dos pesos da rede, foi empregado o algoritmo Adam com taxa de aprendizado inicial de 0.0001. Para evitar o overfitting e melhorar a convergência, foram utilizados dois *callbacks* durante o treinamento. O *ReduceLROnPlateau* que monitora a perda de validação e reduz a taxa de aprendizado em 50% quando não há melhoria por 4 épocas consecutivas e o *ModelCheckpoint* que salva automaticamente apenas o modelo com melhor desempenho na métrica AUC de validação, garantindo que a versão final seja a de maior capacidade de discriminação entre as classes.

4.3. Conjuntos de Dados

Para desenvolver e avaliar o modelo de detecção de glaucoma, este trabalho usou dois conjuntos de dados públicos de imagens de fundoscopia: ACRIMA e RIM-ONE.

O conjunto de dados ACRIMA (*Annotated Retinal Images for Glaucoma Analysis*) foi proposto por Díaz-Pinto et al. (2019) e contém originalmente 705 imagens de fundo de olho, sendo 396 classificadas como glaucoma (Figura 5) e 309 como normais (Figura 6). As imagens foram coletadas na Fundação FISABIO, em Valência, Espanha, com consentimento dos pacientes, e rotuladas por especialistas em glaucoma.

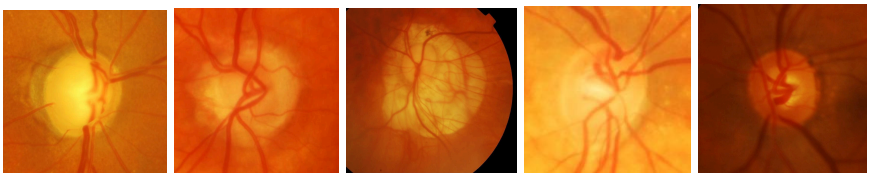

Figura 5. Exemplo de imagens de Glaucoma do conjunto ACRIMA

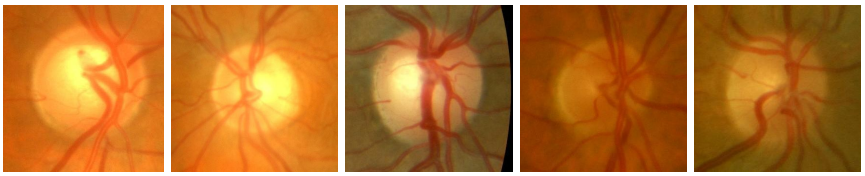

Figura 6. Exemplo de imagens com nervo óptico “normal” do conjunto ACRIMA

Também foi utilizado o conjunto de dados RIM-ONE (*Retinal Image Database for Optic Nerve Evaluation*), originalmente descrito por Fumero et al. (2011). Trata-se de um banco de dados unificado de imagens da retina para avaliação do glaucoma, composto por 172 retinografias de pacientes com glaucoma (Figura 7) e 313

retinografias de indivíduos normais (Figura 8). As imagens foram capturadas em três hospitais espanhóis localizados em Madrid.

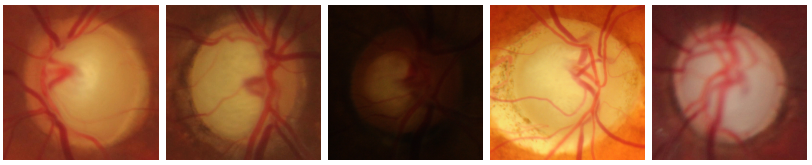

Figura 7. Exemplo de imagens de Glaucoma do conjunto RIM-ONE

Figura 8. Exemplo de imagens com nervo óptico "normal" do conjunto RIM-ONE

4.3.1. Divisão da Base de Dados

Com o *dataset* ACRIMA, para este trabalho, 30 imagens (Glaucoma e Normal) foram selecionadas aleatóriamente para compor o conjunto de testes independente do protótipo. Isso reduziu o total de imagens disponíveis para o treinamento do modelo para 675, das quais 381 são de glaucoma e 294 são normais. Depois de processado, o conjunto foi organizado da seguinte maneira: 471 imagens para treinamento da rede (70%), 102 para validação da rede (15%) e 102 para teste da rede (15%), preservando a proporção original entre as classes em todos os subconjuntos (ver Tabela 1).

Tabela 1. Distribuição do ACRIMA

Distribuição	Glaucoma	Normal	Total de Imagens
Treinamento	266	205	471
Validação	57	45	102
Teste	58	44	102
Teste Protótipo	15	15	30
Total	396	309	705

O RIM-ONE, ao contrário do ACRIMA, exibe um desbalanceamento significativo, refletindo uma situação mais alinhada com a prática clínica real, onde a prevalência de glaucoma é reduzida. Neste estudo, após reservar 30 imagens para avaliação no protótipo, empregaram-se 455 imagens para treinar o modelo, das quais 157 eram de glaucoma e 298 normais. A divisão estratificada gerou 317 imagens para treinamento da rede, 69 para validação da rede e 69 para teste da rede, mantendo a proporção entre as classes (ver Tabela 2).

Tabela 2. Distribuição do RIM-ONE

Distribuição	Glaucoma	Normal	Total de Imagens
Treinamento	109	208	317
Validação	24	45	69
Teste	24	45	69
Teste Protótipo	15	15	30
Total	172	313	485

4.4. Pré-processamento e Data Augmentation

4.4.1. Pré-processamento de Imagens

Todas as imagens dos *datasets* ACRIMA e RIM-ONE foram submetidas a um processo de pré-processamento antes de serem utilizadas no treinamento. Inicialmente, as imagens foram redimensionadas para 224×224 pixels, formato de entrada padrão da arquitetura ResNet50, garantindo compatibilidade com os pesos pré-treinados no ImageNet.

Para melhorar o contraste local e evidenciar estruturas relevantes da retina, como vasos sanguíneos e bordas do disco óptico, foi aplicada a técnica CLAHE (*Contrast Limited Adaptive Histogram Equalization*). Esse recurso opera sobre o canal de luminosidade (L) do espaço de cores CIE L*a*b*, realizando equalização adaptativa de histograma com limite de contraste. Os parâmetros finais utilizados nessa técnica foram *clip limit* (limite de recorte) de 2.0 e tamanho de grade (*tile grid size*) de 8×8 pixels, valores que demonstraram melhor equilíbrio entre realce de detalhes e supressão de ruído nas imagens de fundoscopia.

4.4.2. Data Augmentation

Para aumentar artificialmente a variabilidade do conjunto de treinamento e reduzir o risco de *overfitting*, foi implementada a estratégia de data *augmentation online*, também conhecida como *on-the-fly augmentation*, utilizando a classe ImageDataGenerator do framework Keras. Nesta abordagem, transformações geométricas e fotométricas são aplicadas dinamicamente durante o treinamento, gerando variações aleatórias das imagens a cada época sem necessidade de armazenamento adicional em disco.

As transformações geométricas incluem rotação aleatória de até ±20 graus, deslocamentos horizontais e verticais das dimensões da imagem, cisalhamento de até ±20%, e zoom. Adicionalmente, espelhamentos horizontal e vertical são aplicados aleatoriamente a cada figura durante o treinamento. Quanto às transformações fotométricas, foi aplicada variação de brilho com fator multiplicativo entre 0.7 e 1.3, simulando diferentes condições de iluminação encontradas na prática clínica.

Estas transformações simulam variações naturais encontradas na aquisição de imagens de fundoscopia, como diferentes ângulos de captura, variações na centralização do olho e diferenças entre equipamentos. Diferentemente da *offline augmentation*, onde

novas imagens são geradas e armazenadas previamente, a abordagem online cria variações únicas a cada época de treinamento, resultando em maior diversidade de dados ao longo do processo de aprendizado sem ocupar espaço adicional de armazenamento.

É importante ressaltar que nenhuma transformação de augmentation foi aplicada aos conjuntos de validação e teste, mantendo esses dados em seu estado original para garantir uma avaliação realista e não enviesada do desempenho do modelo. A Tabela 3 apresenta de forma consolidada todos os parâmetros utilizados no pré-processamento e data augmentation.

Tabela 3. Parâmetros utilizados no pré-processamento e Data Augmentation

Etapa	Parâmetro	Valor
CLAHE	Clip limit	2.0
CLAHE	Tamanho do grid	8 × 8
Data augmentation	Rotação	até 20°
Data augmentation	Deslocamento horizontal	até 15%
Data augmentation	Deslocamento vertical	até 15%
Data augmentation	Cisalhamento	0.2
Data augmentation	Zoom	até 10%
Data augmentation	Inversão horizontal	Sim
Data augmentation	Inversão vertical	Sim
Data augmentation	Ajuste de brilho	[0.7, 1.3]

4.5. Desenvolvimento do Protótipo Integrado ao Whatsapp

4.5.1. WhatsApp

Quando uma imagem de retina é recebida de um número previamente autorizado, o sistema identifica automaticamente o tipo de mensagem e verifica se ela se trata de uma mídia válida. Em caso positivo, a imagem é baixada pelo servidor e uma mensagem inicial de feedback é enviada ao profissional da saúde, informando que a análise está em andamento. Em seguida, o módulo estabelece uma conexão com a API de classificação, utilizando um tempo limite de 30 segundos, valor considerado adequado para a execução completa do processo de inferência.

Após o recebimento da resposta da API, o servidor processa os dados retornados e constrói uma mensagem estruturada e de fácil compreensão. Essa resposta inclui o diagnóstico preliminar (Glaucoma ou Normal), o percentual de confiança da classificação, as probabilidades associadas a cada classe e uma ressalva explícita indicando que o resultado corresponde a uma análise automática preliminar, não

substituindo a avaliação de um profissional médico especializado. Por fim, a mensagem é enviada automaticamente ao especialista da saúde, preservando o contexto da conversa.

4.5.2. API de Classificação

O módulo de classificação foi desenvolvido em Python, utilizando o microframework Flask, e disponibiliza um endpoint responsável por receber imagens e retornar diagnósticos. Ao receber uma requisição, a API realiza validações iniciais para garantir a presença do campo de imagem e a existência de um nome de arquivo válido, assegurando a integridade da entrada antes do processamento.

O fluxo de processamento da imagem segue uma sequência bem definida de etapas. Inicialmente, o sistema garante que a imagem esteja no formato RGB, realizando a conversão quando necessário. Em seguida, a imagem é redimensionada para 224×224 pixels, dimensão exigida pela arquitetura ResNet50. Posteriormente, é aplicada a normalização baseada no *dataset* ImageNet, que consiste na subtração da média e divisão pelo desvio padrão de cada canal RGB, conforme valores previamente estabelecidos nesse conjunto de dados.

Com a imagem devidamente pré-processada, o modelo ResNet50 realiza a inferência, retornando uma probabilidade contínua entre 0 e 1, representando a confiança do modelo na presença de glaucoma. Essa probabilidade é interpretada a partir de um limite de decisão fixado em 0,5: valores iguais ou superiores a esse limiar indicam classificação como Glaucoma, enquanto valores inferiores indicam classificação como Normal. A resposta final é estruturada em formato JSON, contendo o diagnóstico textual, a probabilidade bruta retornada pelo modelo, o percentual de confiança e as probabilidades detalhadas associadas a cada classe.

4.6. Avaliação de Desempenho

4.6.1. Métricas de Avaliação do Modelo

Para avaliar o desempenho do modelo ResNet50 na classificação de imagens de glaucoma, foram utilizadas as métricas de acurácia, precisão, sensibilidade (*recall*), especificidade, AUC-ROC, f1-score e matriz de confusão, conforme apresentadas na seção 2.2.3.

A escolha dessas métricas se justifica pela natureza crítica da aplicação em saúde, onde a sensibilidade é fundamental para minimizar falsos negativos (casos de glaucoma não detectados), enquanto a especificidade reduz falsos positivos que geram encaminhamentos desnecessários.

4.6.2. Avaliação Experimental do Protótipo Integrado ao Whatsapp

A avaliação do protótipo teve como propósito analisar o comportamento do sistema integrado em um cenário de uso controlado, observando as respostas retornadas pelo modelo para cada imagem enviada. Diferentemente da avaliação estatística do modelo apresentada na Seção 3.6.1, esta etapa não teve como foco a revalidação do desempenho do classificador, mas sim a análise empírica das respostas geradas pelo protótipo em execução.

Para essa avaliação, foram utilizadas 30 imagens de cada base de dados, sendo 15 pertencentes à classe glaucoma e 15 à classe normal, previamente separadas do processo de treinamento. Cada imagem foi enviada individualmente ao sistema por meio do WhatsApp, permitindo registrar a classe prevista, a probabilidade associada à predição e o percentual de confiança retornado pelo modelo.

A partir dessas respostas, foi possível calcular a taxa de acerto do protótipo e principalmente a confiança média de todas as respostas para cada conjunto avaliado, bem como organizar os resultados em forma de tabelas, possibilitando a comparação entre o comportamento observado no protótipo e os resultados obtidos durante a etapa de avaliação dos modelos no treinamento.

5. Resultados e discussão

5.1. Resultados do Treinamento dos Conjuntos de Dados

A Tabela 4 apresenta os resultados obtidos pelo modelo no conjunto de teste para os *datasets* ACRIMA e RIM-ONE, considerando as métricas de desempenho selecionadas para avaliar sua capacidade de distinguir imagens normais de imagens com glaucoma.

Tabela 4. Métricas de desempenho com os datasets ACRIMA e RIM-ONE

Performance	**ACRIMA**	**RIM-ONE**
Acurácia	98.04%	91.30%
Precisão	98.28%	90.91%
Sensibilidade	98.28%	83.33%
Especificidade	97.73%	95.56%
AUC	99.18%	94.91%
F1 - Score	98.28%	86.96%

A partir dos valores apresentados na Tabela 4, observa-se que o modelo alcançou desempenho elevado no conjunto ACRIMA. A acurácia de 98,04% indica que a maior parte das amostras foi corretamente classificada, refletindo um bom comportamento geral do classificador. Esse resultado é corroborado pela sensibilidade de 98,28%, que evidencia a capacidade do modelo em identificar corretamente imagens de glaucoma, reduzindo de forma significativa a ocorrência de falsos negativos.

A especificidade de 97,73% reforça o desempenho consistente do modelo ao classificar imagens normais, indicando que a taxa de falsos positivos permaneceu baixa. Esse equilíbrio entre sensibilidade e especificidade sugere que o modelo não apenas identifica adequadamente os casos de glaucoma, mas também evita classificações incorretas de indivíduos saudáveis, o que é desejável em sistemas de apoio à decisão.

A Figura 9 apresenta a curva ROC correspondente ao conjunto ACRIMA. Verifica-se que a curva se mantém distante da diagonal aleatória ao longo de praticamente todo o intervalo, com rápida elevação da taxa de verdadeiros positivos

mesmo para valores reduzidos da taxa de falsos positivos. O valor de AUC de aproximadamente 0,992 confirma a elevada capacidade discriminativa do modelo, indicando que as probabilidades estimadas permitem uma separação consistente entre as classes normal e glaucoma em diferentes limites de decisão.

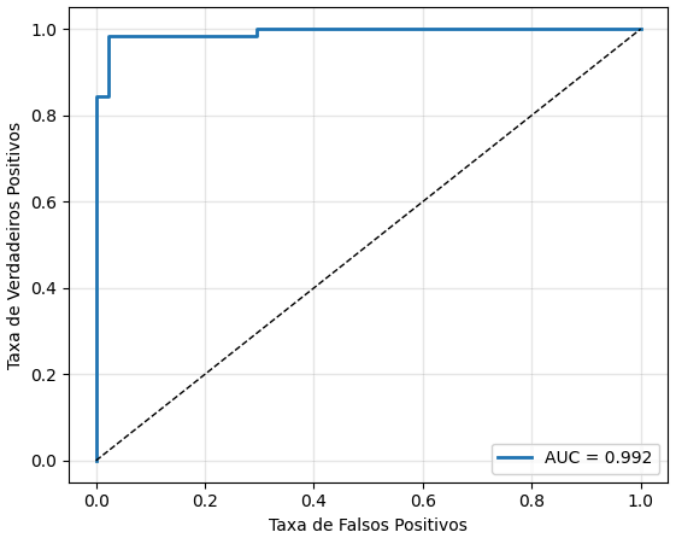

Figura 9. Curva ROC ACRIMA

A Figura 10 apresenta a matriz de confusão obtida para o conjunto de testes da base de dados ACRIMA. A análise dos resultados evidencia um desempenho consistente do modelo, com baixos índices de erro em ambas as classes. Das 44 imagens normais avaliadas, 43 foram corretamente classificadas, enquanto apenas uma foi incorretamente identificada como glaucoma. Esse resultado confirma a elevada capacidade do modelo em reconhecer imagens sem a presença da doença, mantendo uma taxa reduzida de falsos positivos e reforçando o valor da especificidade observada anteriormente.

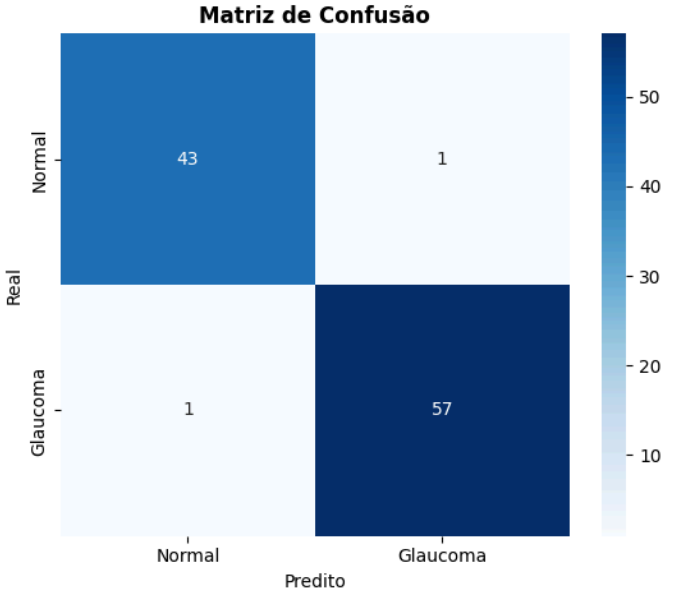

Figura 10. Matriz de Confusão ACRIMA

No que se refere às imagens com glaucoma, o modelo classificou corretamente 57 das 58 amostras presentes no conjunto de teste, ocorrendo apenas um falso negativo. Esse resultado evidencia uma elevada sensibilidade, indicando que o modelo apresentou grande eficiência na identificação de casos positivos de glaucoma, com mínima perda de amostras relevantes.

De forma geral, a matriz de confusão do conjunto ACRIMA demonstra um equilíbrio favorável entre a identificação correta de imagens normais e glaucomatosas, com número reduzido de erros em ambas as classes. Esse comportamento reforça a robustez do modelo nesse conjunto de dados e está alinhado com os elevados valores de acurácia e AUC previamente apresentados, indicando uma separação clara entre as classes no espaço de decisão do classificador.

No conjunto RIM-ONE, o modelo também apresentou resultados satisfatórios, ainda que inferiores aos observados no conjunto ACRIMA. A acurácia de 91,30% indica um bom desempenho geral na classificação das imagens, demonstrando que a maior parte das amostras foi corretamente identificada. Entretanto, nota-se uma redução mais expressiva na sensibilidade, que atingiu 83,33%, evidenciando maior dificuldade do modelo em identificar todos os casos positivos de glaucoma nesse conjunto de dados.

Em contrapartida, a especificidade manteve-se elevada, com valor de 95,56%, o que indica que o modelo preservou uma boa capacidade de reconhecer corretamente imagens normais. Esse comportamento sugere que, embora o modelo tenha apresentado maior incidência de falsos negativos no RIM-ONE, ele permaneceu consistente na identificação de indivíduos sem a doença, reduzindo a ocorrência de falsos positivos.

A curva ROC correspondente ao conjunto RIM-ONE é apresentada na Figura 11. Observa-se que, apesar de a curva permanecer acima da diagonal aleatória ao longo de todo o intervalo, sua inclinação inicial é menos acentuada quando comparada à obtida no conjunto ACRIMA. O valor de AUC de 0,9491 confirma uma boa capacidade discriminativa do modelo, embora indique maior sobreposição entre as distribuições das classes, o que pode estar associado às características intrínsecas do conjunto de dados.

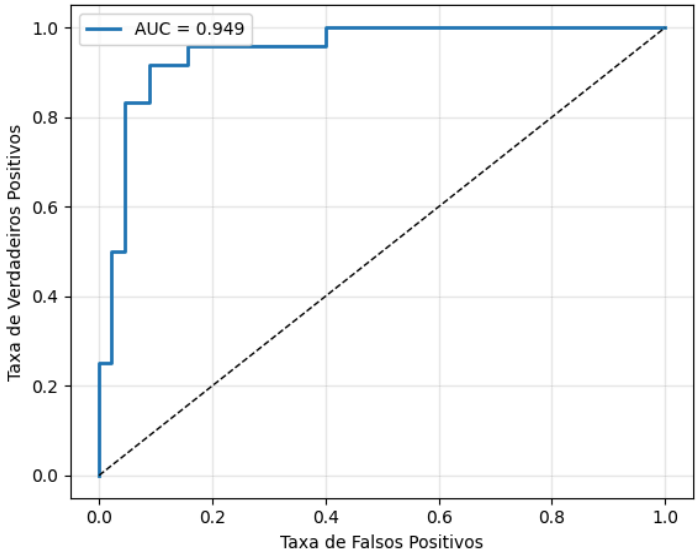

Figura 11. Curva ROC RIM-ONE

A figura 12 apresenta a matriz de confusão obtida para o conjunto de teste do *dataset* RIM-ONE, considerando a configuração experimental adotada. A partir dessa matriz, é possível analisar de forma detalhada os acertos e erros cometidos pelo modelo na classificação das imagens.

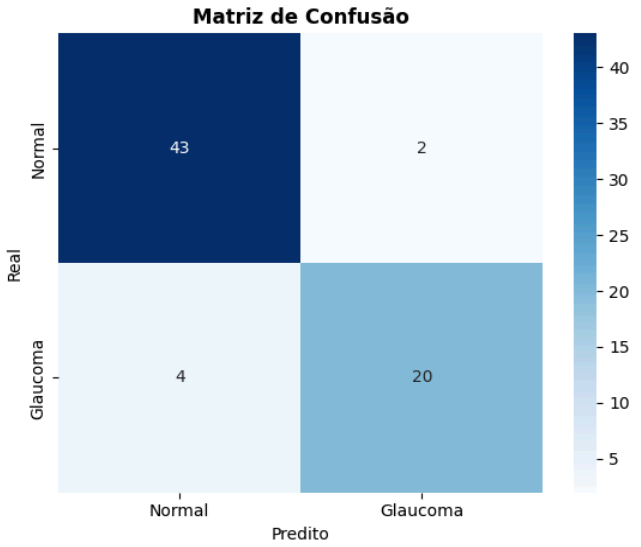

Figura 12. Matriz de Confusão RIM-ONE

Observa-se que, das 45 imagens normais presentes no conjunto de teste, 43 foram corretamente classificadas como normais, enquanto apenas 2 foram incorretamente identificadas como glaucoma. Esse resultado evidencia a elevada especificidade do modelo nesse conjunto, confirmando sua capacidade de reconhecer adequadamente imagens sem a presença da doença e mantendo uma baixa taxa de falsos positivos.

Em relação às imagens glaucomatosas, das 24 amostras avaliadas, 20 foram corretamente classificadas, enquanto 4 foram classificadas de forma incorreta como normais. Esses erros correspondem aos falsos negativos e estão diretamente associados à redução observada na sensibilidade do modelo para o conjunto RIM-ONE. Em um cenário de triagem, esse tipo de erro é particularmente relevante, pois representa casos em que a doença pode não ser detectada pelo sistema automatizado.

De modo geral, a matriz de confusão confirma os resultados apresentados anteriormente pelas métricas quantitativas na Tabela 4. O modelo demonstra um comportamento mais conservador, com maior facilidade em identificar imagens normais do que em detectar todos os casos de glaucoma nesse conjunto de dados. Esse padrão sugere que características particulares da base RIM-ONE, como variabilidade visual entre as imagens ou maior sobreposição entre as classes, influenciam diretamente o desempenho do classificador.

5.2. Resultados da Experimentação com o Protótipo Integrado ao Whatsapp

5.2.1. Resultados da Avaliação do Protótipo com o Dataset ACRIMA

No conjunto de imagens da classe Glaucoma, o protótipo obteve 13 acertos em 15 amostras, correspondendo a uma taxa de acerto de 86,67%. Observou-se que a maior parte das predições corretas apresentou níveis elevados de confiança, frequentemente acima de 90%. Conforme é possível constatar na Figura 13 que traz o resultado de uma das análises do protótipo, tendo 99.76% de confiança sobre Glaucoma na sua resposta. Os dois casos classificados incorretamente ocorreram em imagens nas quais a probabilidade atribuída à classe Normal foi superior à de Glaucoma, indicando maior ambiguidade visual nesses exemplos.

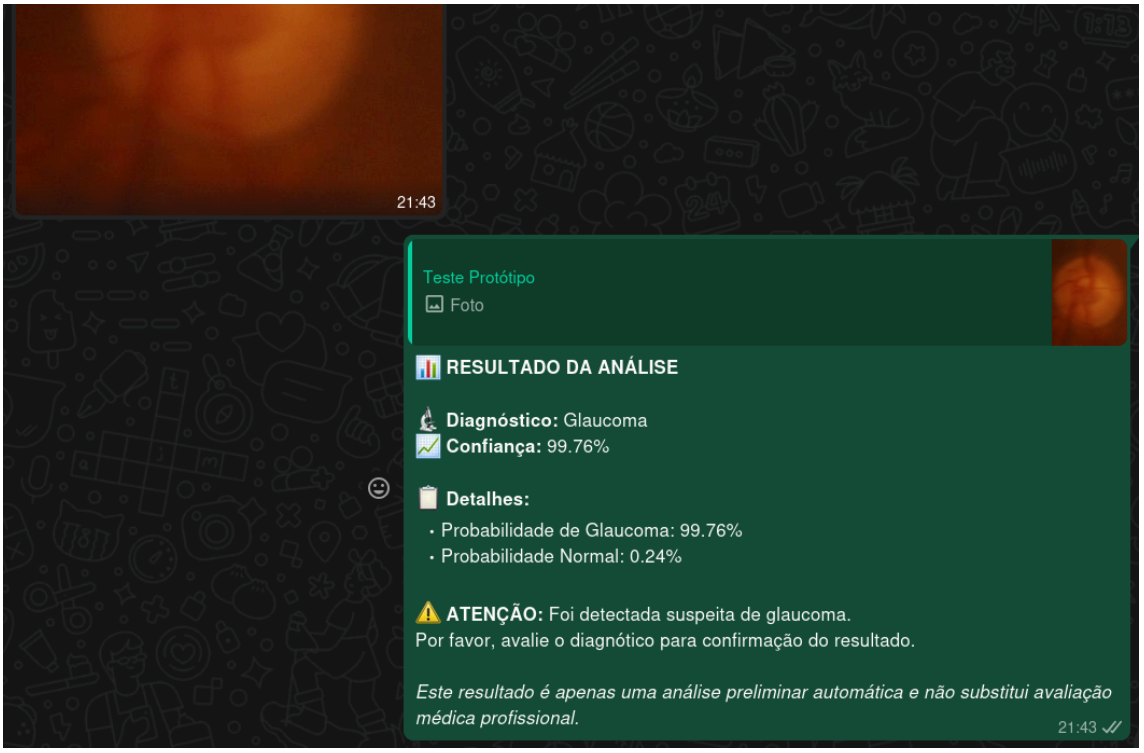

Figura 13. Demonstração do protótipo para classe "Glaucoma" do ACRIMA

Para o conjunto de imagens da classe Normal, o protótipo obteve 12 acertos em 15 amostras, resultando em uma taxa de acerto de 80%. Os três erros observados correspondem a imagens classificadas como Glaucoma com níveis de confiança variando entre moderados e elevados, indicando que determinadas amostras normais apresentam características visuais que podem induzir o modelo a uma interpretação equivocada.

Conforme apresentado na Tabela 5, o protótipo obteve maior confiança média nas classificações da classe Glaucoma (93,8%) em comparação à classe Normal (91.73%). Isso se deu a partir da média de cada categoria (Glaucoma e Normal). Apesar da ocorrência de erros em ambas as classes, observa-se que, mesmo nos casos incorretos, o modelo frequentemente apresentou níveis elevados de confiança, especialmente para imagens normais classificadas como Glaucoma. Neste sentido, cinco amostras de imagens do tipo normal tiveram 100% de probabilidade no seu diagnóstico,

resultados obtidos somente com esse banco de dados e classe, conforme é possível observar a demonstração de resposta de diagnóstico na Figura 14.

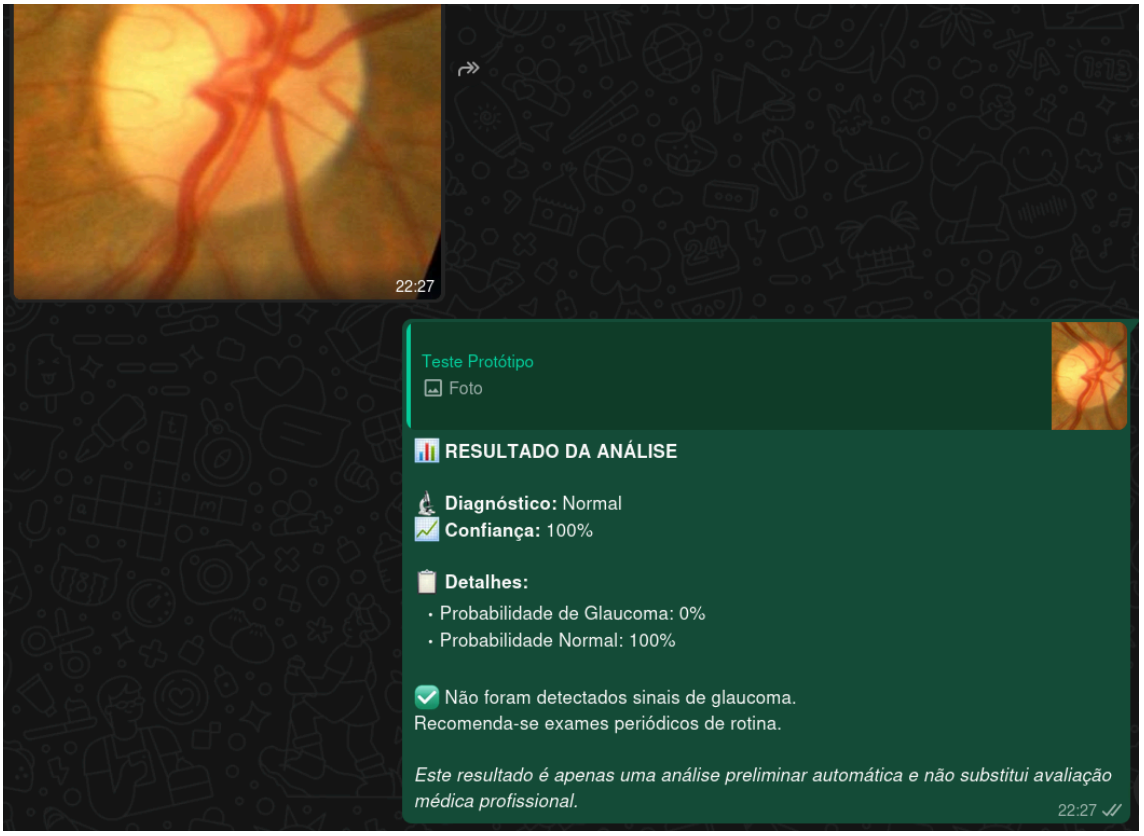

Figura 14. Demonstração do protótipo para classe "Normal" com ACRIMA

Esse comportamento indica a presença de padrões visuais semelhantes entre algumas amostras, reforçando a necessidade de interpretação cautelosa dos resultados do protótipo.

Tabela 5. Resultados da avaliação do protótipo com imagens do dataset ACRIMA

Classe Real	Total	Acertos	Erros	Confiança por classe (%)
Glaucoma	15	13	2	97.62%
Normal	15	12	3	91.73%
Geral	30	25	5	94.80%

De forma geral, os resultados evidenciam que o protótipo apresentou desempenho superior na identificação de imagens com glaucoma em comparação à classe Normal, comportamento que está em consonância com o foco do modelo em priorizar a detecção de casos positivos.

5.2.2. Resultados da Avaliação do Protótipo com o Dataset RIM-ONE

Dos resultados obtidos com a experimentação do protótipo utilizando as imagens do conjunto RIM-ONE, é apresentado na Tabela 6 que no conjunto de imagens da classe Glaucoma foram observados 11 acertos e 4 erros, resultando em uma taxa de acerto de

73,3%. A confiança média associada à todas as classificações corretas dessa classe foi de 85,66%, onde é possível visualizar na Figura 15 uma amostra de resposta obtida pelo protótipo em que ela atingiu 95.08% de confiança/probabilidade na sua análise.

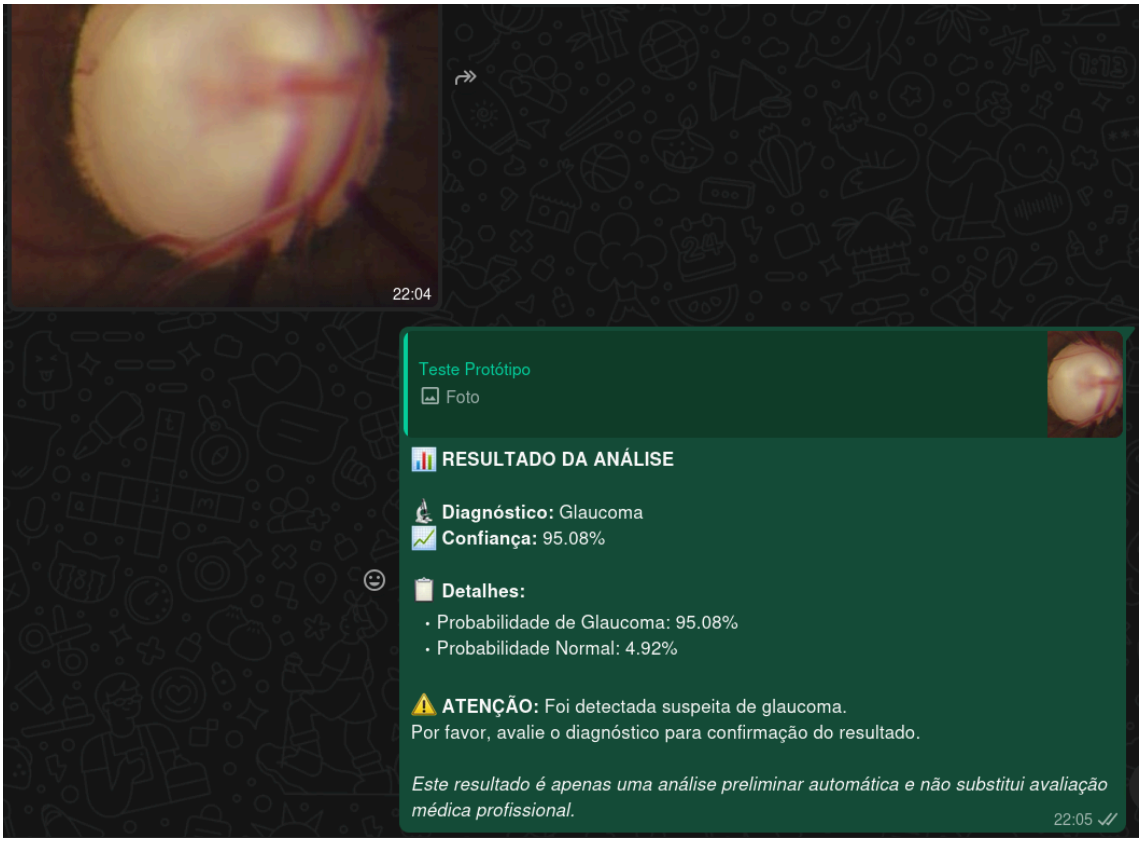

Figura 15. Demonstração do protótipo para classe "Glaucoma" com RIM-ONE

No entanto, para as imagens da classe Normal, ainda na Tabela 6 pode-se visualizar que o protótipo apresentou 13 classificações corretas de um total de 15 imagens, correspondendo a uma taxa de acerto de 86,7%, com 2 classificações incorretas. A confiança média de todas as predições corretas dessa classe foi de 88,74%, indicando que, na maioria dos casos, o modelo atribui alta probabilidade à classe correta, que nesse caso é Normal.

Tabela 6. Resultados da avaliação do protótipo com imagens do dataset RIM-ONE

Classe Real	Total	Acertos	Erros	Confiança por classe (%)
Glaucoma	15	11	4	85.66%
Normal	15	13	2	88.74%
Geral	30	24	6	87.31%

Considerando ambas as classes, o protótipo alcançou 24 acertos em 30 imagens, o que corresponde a uma taxa global de acerto de 80%. A confiança média geral das classificações corretas foi de 87,31%, valor que reforça a consistência das respostas fornecidas pelo sistema durante a experimentação.

No caso do RIM-ONE, observou-se maior oscilação no comportamento das predições, tanto nas taxas de acerto quanto nos níveis médios de confiança. Ainda que a confiança média tenha permanecido elevada nas classificações corretas, a ocorrência de erros com probabilidades significativas indica maior sensibilidade do modelo à variabilidade das imagens presentes nesse conjunto. Esse aspecto evidencia que a diversidade visual pode impactar a estabilidade das decisões do sistema. Do ponto de vista prático, os resultados confirmam que o protótipo mantém desempenho consistente para fins de triagem, porém reforçam a necessidade de validações adicionais em bases mais heterogêneas antes de qualquer aplicação em contexto clínico real.

5.3. Comparação dos Resultados

Ao comparar os resultados do modelo para cada *dataset*, foi possível observar que o conjunto ACRIMA foi o que apresentou resultados mais satisfatórios, tendo como base a acurácia de 98,04% durante o treinamento e taxa de confiança média de todas as respostas de 94.80% na experimentação com o protótipo. Em contrapartida, o *dataset* RIM-ONE apresentou desempenho inferior, com acurácia de 91,30% no treinamento e 87.31% no protótipo, evidenciando maior dificuldade do modelo em generalizar para este conjunto de dados.

A diferença de desempenho entre os dois conjuntos de dados pode ser explicada por suas características particulares. No caso do ACRIMA em que apresenta distribuição relativamente equilibrada entre as classes, com 56,5% de imagens com glaucoma e 43,5% de imagens normais, enquanto o RIM-ONE possui desbalanceamento mais acentuado, sendo 34,4% de glaucoma e 65,6% de imagens normais. Esse desbalanceamento reflete diretamente na sensibilidade do modelo, que alcançou 98,28% no ACRIMA, mas reduziu para 83,33% no RIM-ONE.

Em relação à confiança das predições, observou-se que o protótipo apresentou confiança média superior no conjunto ACRIMA, com valores de 97.62% para a classe Glaucoma e 91.73% para a classe Normal, em comparação aos 85,66% e 88,74% observados no RIM-ONE, respectivamente. Essa conduta sugere que o modelo identifica padrões mais distintivos no *dataset* ACRIMA, resultando em maior certeza nas classificações. Entretanto, mesmo em casos de erro, o protótipo frequentemente apresentou níveis elevados de confiança, fenômeno conhecido como overconfidence, comum em redes neurais profundas quando aplicadas em bancos de dados de tamanho limitado. Esse aspecto é particularmente relevante para uso clínico, pois predições incorretas com alta confiança podem induzir interpretações equivocadas por parte de profissionais de saúde que utilizam o sistema como ferramenta de apoio.

De forma geral, os resultados demonstram que o modelo apresentou desempenho consistentemente superior no conjunto de dados ACRIMA em todas as métricas avaliadas, tanto durante o treinamento quanto na experimentação prática com o protótipo. A diferença de desempenho entre os dois conjuntos de dados evidencia a influência de fatores como distribuição de classes, variabilidade visual das imagens e características específicas de aquisição dos dados.

6. Considerações Finais

Este trabalho desenvolveu e avaliou um sistema automatizado de detecção de glaucoma em imagens de fundoscopia utilizando a arquitetura ResNet50 com transfer learning, integrando-o a um protótipo funcional com interação via WhatsApp. Os resultados demonstraram desempenho superior no dataset ACRIMA, alcançando acurácia de 98,04%, precisão de 98,28% e AUC de 99,18%, em comparação ao RIM-ONE, evidenciando a influência direta da qualidade, padronização e do balanceamento dos dados no desempenho de modelos baseados em aprendizado profundo. Esses resultados mostram-se competitivos em relação à literatura, como observado no estudo de Sreng et al. (2020), que reportou precisão de 99,53% e AUC de 99,98% utilizando o mesmo dataset.

A implementação do protótipo permitiu avaliar o modelo em um cenário mais próximo do uso real. Embora tenha sido observada leve redução de desempenho em relação à fase de treinamento, a integração mostrou-se tecnicamente viável. Entre as principais restrições do estudo, destacam-se o tamanho limitado dos datasets e as limitações computacionais, que impediram a exploração de arquiteturas mais complexas e a validação em bases clínicas reais. Ainda assim, os resultados obtidos confirmam o potencial do uso de redes neurais convolucionais como ferramenta auxiliar na triagem de glaucoma.

Como trabalhos futuros, sugere-se a ampliação da validação para outros datasets e imagens adquiridas em ambiente clínico real, bem como a incorporação de técnicas de interpretabilidade, como o Grad-CAM (Gradient-weighted Class Activation Mapping), a fim de evidenciar as regiões da imagem que influenciam as decisões do modelo e aumentar a confiança na aplicação clínica.

Conclui-se que a integração entre modelos de deep learning e plataformas amplamente utilizadas de comunicação pode ampliar o acesso a ferramentas de apoio diagnóstico, contribuindo para o rastreamento precoce do glaucoma. Ressalta-se, contudo, que tais sistemas devem atuar exclusivamente como suporte à decisão clínica, não substituindo a avaliação médica especializada.

Referências

Baudouin, C.; Kolko, M.; Melik-Parsadaniantz, S.; Messmer, E. M. (2021). Inflammation in glaucoma: From the back to the front of the eye, and beyond. Progress in Retinal and Eye Research, 83, 100916. https://doi.org/10.1016/j.preteyeres.2020.100916

Bragança, C. P.; Torres, J.M.; Soares, C.P.d.A.; 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. *Healthcare*, 10, 2345.

Braz, G. P. et al. (2025). Modelos de triagem em serviços de urgência: impactos na qualidade assistencial e tempo de resposta. *Cognitus Interdisciplinary Journal*, 2(3), 147–160.

Christopher, M.; Belghith, A.; Bowd, C.; Proudfoot, J. A.; Goldbaum, M. H.; Weinreb, R. N.; Girkin, C. A.; Liebmann, J. M.; Zangwill, L. M. (2018). Performance of deep

learning architectures and transfer learning for detecting glaucomatous optic neuropathy in fundus photographs. *Scientific Reports*, 8, 16685.

Das Graças Mendes Júnior, L.; De Lima, G. M.; Oliveira Tavares, J.; Mendonça, L. R. A.; Silva Narciso, T. (2025) O uso da inteligência artificial no auxílio da triagem de adultos em serviços de emergência. RECIMA21 – Revista Científica Multidisciplinar, v. 6, n. 10, e6106821. DOI: https://doi.org/10.47820/recima21.v6i10.6821.

Díaz-Pinto, A., Morales, S., Naranjo, V., Köhler, T., Mossi, J. M., and Navea, A. (2019). CNNs for automatic glaucoma assessment using fundus images: an extensive validation. Biomedical Engineering Online, 18(1), 1–19.

Fawcett, T. (2006). An introduction to ROC analysis. *Pattern Recognition Letters*, 27(8), 861–874.

Fumero, F., Alayón, S., Sanchez, J. L., Sigut, J., and González-Hernández, M. (2011). RIM-ONE: An open retinal image database for optic nerve evaluation. In: *Proceedings of the 24th International Symposium on Computer-Based Medical Systems* (CBMS), pp. 1–6.

Gomes, J. S.; Brito, L. E. S.; Dalília, A.; Veloso, R. R.; Carvalho, A. O.; Araújo, F. H. (2025). Detecção automática do glaucoma em imagens retinianas utilizando redes neurais convolucionais. In: *Simpósio Brasileiro de Computação Aplicada à Saúde* (SBCAS). Anais. Porto Alegre: SBC, p. 629–640.

Goodfellow*, I.;* Bengio, Y; Courville*, A. (2016) Deep Learning. Cambridge, MA:* MIT Press. ISBN 9780262337373.

Gulshan, V.; Peng, L.; Coram, M.; Stumpe, M. C.; Wu, D.; Narayanaswamy, A.; Venugopalan, S.; Widner, K.; Madams, T.; Cuadros, J.; Kim, R.; Raman, R.; Nelson, P. C.; Mega, J. L.; Webster, D. R. (2016). Development and Validation of a Deep Learning Algorithm for Detection of Diabetic Retinopathy in Retinal Fundus Photographs. *JAMA*, 316(22), 2402–2410. https://doi.org/10.1001/jama.2016.17216.

He, K.; Zhang, X.; Ren, S.; Sun, J. (2016). Deep residual learning for image recognition. In: *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition* (CVPR), p. 770–778. https://doi.org/10.1109/CVPR.2016.90.

LeCun, Y.; Bengio, Y. (1995) Convolutional Networks for Images, Speech, and Time-Series. The Handbook of Brain Theory and Neural Networks. MIT Press, pp. 255–258.

Li, Z.; He, Y.; Keel, S.; Meng, W.; Chang, R. T.; He, M. (2018). Efficacy of a Deep Learning System for Detecting Glaucomatous Optic Neuropathy Based on Color Fundus Photographs. *Ophthalmology*, 125(8), 1199–1206.

Litjens, G.; Kooi, T.; Bejnordi, B. E.; Setio, A. A. A.; Ciompi, F.; Ghafoorian, M.; van der Laak, J. A. W. M.; van Ginneken, B.; Sánchez, C. I. (2017). A survey on deep learning in medical image analysis. *Medical Image Analysis*, 42, 60–88.

Mars, M. and Scott, R. E. (2017). WhatsApp in Clinical Practice: A Literature Review. *Studies in Health Technology and Informatics*, v. 231, p. 82-90.

Neomed. (2025). Inteligência artificial na medicina: 5 cases de aplicações bem sucedidas. Disponível em: https://neomed.com.br/inteligencia-artificial-na-medicina-conheca-5-cases-de-aplicacoes-bem-sucedidas/. Acesso em: 15 dez. 2025.

Paiva, A. M. M.; Gomes, V. S.; Takaoka, R. A. (2025). Inteligência artificial na triagem de urgência e emergência: o papel da enfermagem na otimização da assistência. *Revista FOCO*, 18(12), e10979.

Powers, D. M. W. (2011). Evaluation: From precision, recall and F-measure to ROC, informedness, markedness and correlation. *Journal of Machine Learning Technologies*, 2(1), 37–63.

Prodanov, C. C.; Freitas, E. C. de. (2013). *Metodologia do trabalho científico: métodos e técnicas da pesquisa e do trabalho acadêmico*. 2. ed. Novo Hamburgo: Feevale.

Rezende, F. A.; Coelho, F. A.; Lima, V. C. (2020). A importância do diagnóstico precoce nas doenças oculares. *Arquivos Brasileiros de Oftalmologia*, 83(3), 205–212.

Sreng, S.; Maneerat, N.; Hamamoto, K.; Win, K. Y.; (2020). Deep learning for optic disc segmentation and glaucoma diagnosis on retinal images, *Applied Sciences*, vol. 10, no. 14, pp. 1–19.

Shibata, N.; Tanito, M.; Mitsuhashi, K.; Fujino, Y.; Matsuura, M.; Murata, H.; Asaoka, R. (2018). Development of a deep residual learning algorithm to screen for glaucoma from fundus photography. *Scientific Reports*, 8, 14665.

Tamim, N., Elshrkawey, M., and Nassar, H. (2021). Accurate diagnosis of diabetic retinopathy and glaucoma using retinal fundus images based on hybrid features and genetic algorithm. *Applied Sciences*, 11(13):6178.

Tham Y. C, Li X., Wong T. Y., Quigley H. A., Aung T., Cheng C. Y. (2014). Global prevalence of glaucoma and projections of glaucoma burden through 2040: a systematic review and meta-analysis. *Ophthalmology*. 121(11):2081-90.

Ting, D. S. W.; Cheung, C. Y.; Lim, G.; Tan, G. S. W.; Quang, N. D.; Gan, A.; Hamzah, H.; Garcia-Franco, R.; San Yeo, I. Y.; Lee, S. Y.; Wong, E. Y. M.; Sabanayagam, C.; (2017). Development and Validation of a Deep Learning System for Diabetic Retinopathy and Related Eye Diseases Using Retinal Images From Multiethnic Populations With Diabetes. *JAMA*, 318(22), 2211–2223.

Voulodimos, A.; Doulamis, N.; Doulamis, A.; Protopapadakis, E. (2018). Deep learning for computer vision: a brief review. *Computational Intelligence and Neuroscience*, 2018, 7068349.

Weinreb, R. N.; Aung, T.; Medeiros, F. A. (2014). The pathophysiology and treatment of glaucoma. *JAMA*, 311(18), 1901–1911.

World Health Organization (WHO). (2019). World report on vision. Geneva: WHO.

Xavier, P. B. (2024). A utilização das tecnologias digitais na assistência em saúde. *Revista Eletrônica Acervo Saúde*, v. 24, n. 4, p. e16136-e16136.