UNIVERSIDADE FEDERAL DO MARANHÃO

Curso de Ciência da Computação

Lucas Cunha de Carvalho

Estudo Comparativo de Arquiteturas de Redes Neurais Convolucionais aplicado ao Diagnóstico de Patologias da Visão

São Luís

2023

Lucas Cunha de Carvalho

Estudo Comparativo de Arquiteturas de Redes Neurais Convolucionais aplicado ao Diagnóstico de Patologias da Visão

Monografia apresentada ao curso de Ciência da Computação da Universidade Federal do Maranhão, como parte dos requisitos necessários para obtenção do grau de Bacharel em Ciência da Computação.

Orientador: Prof. Dr. Geraldo Braz Júnior

São Luís

2023

Ficha gerada por meio do SIGAA/Biblioteca com dados fornecidos pelo(a) autor(a).
Diretoria Integrada de Bibliotecas/UFMA

Carvalho, Lucas Cunha de.
Estudo Comparativo de Arquiteturas de Redes Neurais Convolucionais aplicado ao Diagnóstico de Patologias da Visão / Lucas Cunha de Carvalho. - 2023.
40 p.

Orientador(a): Geraldo Braz Junior.
Monografia (Graduação) - Curso de Ciência da Computação, Universidade Federal do Maranhão, São Luís - Maranhão, 2023.

1. Aprendizado profundo. 2. ConvNeXt. 3. EfficientNet. 4. Imagem de fundo de olho. I. Junior, Geraldo Braz. II. Título.

Lucas Cunha de Carvalho

Estudo Comparativo de Arquiteturas de Redes Neurais Convolucionais aplicado ao Diagnóstico de Patologias da Visão

Monografia apresentada ao curso de Ciência da Computação da Universidade Federal do Maranhão, como parte dos requisitos necessários para obtenção do grau de Bacharel em Ciência da Computação.

Trabalho aprovado em São Luís, 11 de janeiro de 2023:

Prof. Dr. Geraldo Braz Júnior
Orientador

Prof. Dra. Simara Vieira da Rocha
Examinador

Profa. Dr. Anselmo Cardoso de Paiva
Examinador

São Luís
2023

Agradecimentos

Primeiramente à minha família, em especial à minha mãe Rita que sempre apoiou e se preocupou com a minha educação, para ela minha saúde e aprendizado eram prioridades, não seria o que sou hoje se não fosse por ela. Agradeço à minha tia Rosana e avó Josefina que são verdadeiros pilares na minha vida e a meu pai Sidney que também sempre apoiou minha educação.

Agradeço ao DEINF e seus professores, em especial, meu orientador Geraldo, pelo qual declaro minha profunda gratidão e admiração, desde que o conheci na disciplina de linguagem de programação têm sido uma referência e um grande mentor na área e que profissionalmente abriu portas para mim com a oportunidade de participar do PIBIC em 2020.

Também agradeço a todos os professores da Universidade Federal do Maranhão, com suas aulas inesquecíveis e que mesmo nas dificuldades com reprovações, pude aprender e tornar-me um homem muito mais forte e capaz do que quando entrei em 2017, estes pouco mais de 5 anos foram de muito aprendizado e sobretudo de amadurecimento pessoal.

Por fim agradeço à meus amigos que suavizaram toda a jornada com seu carinho e companheirismo, os que conheci na UFMA: Italo, Hanna, Michele e Renata são amizades que carregarei para sempre, á Hanna e meu amigo de longa data Cristian e sua família, meu agradecimento especial por terem sido uma verdadeira terapia para mim nesses anos de graduação.

"Em homenagem à minha amada mãe.
Que sempre me apoiou e dedicou sua vida à mim.
Razão maior de meu sucesso.
Te amo para todo o sempre."

Resumo

O diagnóstico precoce de patologias como a retinopatia diabética e glaucoma é fundamental para o tratamento por conta do risco de cegueira, entretanto, essa demanda muitas vezes não pode ser atendida devido ao processo de análise manual que costuma ser demorado além da grande quantidade de exames realizados visando detectar tais patologias que crescem cada vez mais em todo o mundo. Técnicas de aprendizagem profunda podem ser fundamentais na busca pela automatização de tal processo de diagnose de patologias relacionadas ao fundo de olho, por meio da detecção entre imagens patológicas ou saudáveis, favorecendo o diagnóstico precoce, a agilidade em fornecer resultados de novos exames, além de reduzir custos e esforços humanos. Para um trabalho de classificação com tantas características a serem analisadas nas imagens, a escolha da rede neural ideal é fundamental para melhores resultados e economia de custos computacionais. Este estudo realiza uma análise entre duas redes neurais de convolução que possuem abordagens diferentes, a ConvNeXt e EfficientNet. Estas baseam seus funcionamentos em respectivamente, Vision Transformers e escalabilidade de parâmetros em uma rede de convolução convencional. O trabalho proposto também pôde analisar técnicas de pré-processamento de imagens, otimização de parâmetros e balanceamento de imagens, além de testar e avaliar diferentes métodos de treinamento como Shallow Fine Tuning e Deep Fine Tuning sendo os experimentos efetuados em 5748 imagens de fundo de olho da base *ODIR-5K*. Através das técnicas citadas, os modelos obtiveram médias de métricas entre *Precision*, *Recall* e *F1-Score* de 0,71 para a ConvNeXt e 0,78 para a EfficientNet, resultando em modelos que puderam aprender características das imagens e que foram capazes de detectar imagens de fundo de olho patológicas ou saudáveis com precisão satisfatória.

Palavras-chave: ConvNeXt, EfficientNet, Aprendizado profundo, Imagem de fundo de olho.

Abstract

Early diagnosis of pathologies such as diabetic retinopathy and glaucoma is essential for treatment due to the risk of blindness, however, this demand often not answered due to the manual review process which is often time consuming in addition to the large number of tests carried out to detect such pathologies that are growing all over the world. Deep learning techniques can be fundamental in the search for the automation of such process of diagnosis of pathologies related to the fundus of the eye, through the detection between pathological or healthy images, favoring early diagnosis, agility in providing results of new exams, in addition to reducing costs and human efforts. For a classification problem with so many features in the images, the choice of the ideal neural network is essential for better results and computational cost savings, this study performs an analysis between two convolution neural networks that have different approaches, the ConvNeXt and EfficientNet, which base their functioning on respectively, Vision Transformers and parameter scalability in a conventional convolution network. The proposed work could also analyze image pre-processing, parameter optimization and image balancing techniques, as well as testing and evaluating different methods of training such as Shallow Fine Tuning and Deep Fine Tuning the experiments were performed on 5748 fundus images from the *ODIR-5K* base. Through the techniques cited above, the models obtained mean metrics between *Precision*, *Recall* and *F1-Score* of 0.71 for ConvNeXt and 0.78 for EfficientNet, resulting in models that could learn image features and who were able to detect pathological or healthy fundus images with satisfactory accuracy.

Keywords: ConvNeXt, EfficientNet, Deep Learning, Fundus eye images.

Lista de ilustrações

Lista de tabelas

Lista de abreviaturas e siglas

CNN	*Convolution Neural Network (Rede Neural Convolucional)*
ViT	*Vision Transformer*
SwimT	*Swim Transformer*
FLOPS	*Floating point operations per second (Operações de ponto flutuante por segundo)*

Sumário

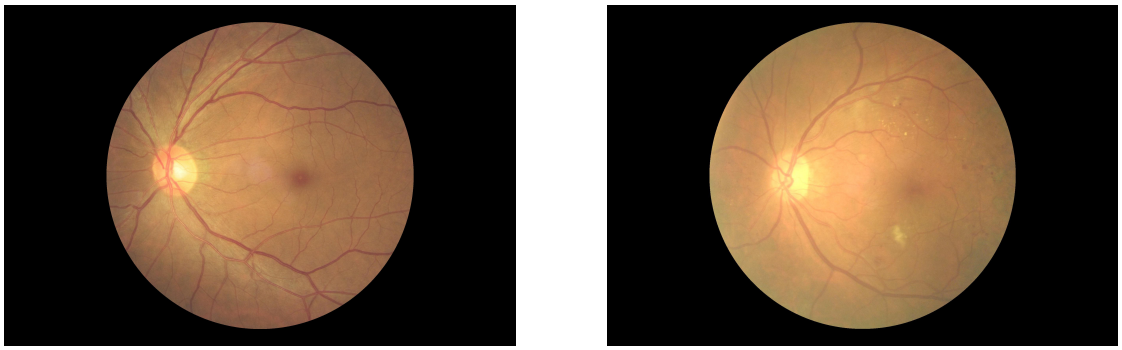
1 Introdução

Doenças relacionadas á visão como a retinopatia diabética e o glaucoma afetam pessoas por todo o mundo. A detecção precoce de tais patologias se faz necessária para prognósticos positivos, entretanto, tal trabalho demanda tempo, esforço intensivo e profissionais capacitados. Tais variáveis representam desafios ao diagnóstico. Esta problemática se agrava à medida em que as prevalências destas doenças crescem de forma contínua ao longo dos anos, consequentemente, aumentando a demanda por serviços de triagem.

Tradicionalmente, imagens de fundo de olho são utilizadas para detectar e acompanhar a evolução de patologias, para tal, faz-se necessário o reconhecimento de diferentes padrões formados que possam identificar o estado de saúde do órgão, tais padrões podem ser segmentos dos olhos que por vezes possuem lesões e alterações que configuram informações necessárias ao disgnóstico, a automação do processo de identificação de um olho patológico utilizando técnicas de deep learning se mostra uma ferramenta essencial ao auxílio da detecção precoce de patologias.

A identificação e classificação eficiente e eficaz de olhos sintomáticos é um trabalho complicado devido às grandes variações na massa, tonalidade, orientação e formas das lesões (Figura 1). Além disso, a extensa semelhança entre lesões e estruturas naturais e entre as cores dos olhos complicam ainda mais o processo de classificação, em diversos exemplos, uma imagem de fundo de olho considerada patológica pode ser confundida como saudável se uma análise descuidada for efetuada.

Em uma problemática com tantas variáveis, poderosas ferramentas de inteligência artificial como as redes neurais podem ser utilizadas para aprender dos mais diversos padrões e buscar identificar com mais eficiência a presença ou não de algum tipo de patologia



Olho saudável. Olho Diabético.

Figura 1 – Imagens de fundo de olho

em uma imagem. É importante que seja escolhido de forma criteriosa a arquitetura de rede neural ideal para o problema especificado.

Na escolha da rede neural ideal, é fundamental a análise das mais diversas arquiteturas, entre elas, as estado da arte, nos últimos tempos, uma arquitetura têm tomado o destaque entre outros modelos, a EfficientNet Tan e Le (2019), uma arquitetura de rede neural convolucional altamente otimizada, criada com o propósito da resolução de problemas de classificação e que busca um melhor desempenho, equilibrando profundidade, largura e resolução da rede.

Tal arquitetura é capaz de obter altas precisões mantendo um desempenho poderoso,com baixo custo computacional e alta velocidade de execução no aprendizado dos mais diversos problemas de deep learning.

Outro conceito que têm ganhado destaque e espaço no que se refere a problemas que exigem aprendizado profundo é o de transformers. Desde a sua introdução, a arquitetura Transformer tornou-se a arquitetura dominante em tarefas de processamento de linguagem natural, substituindo arquiteturas recorrentes anteriormente populares Touvron et al. (2022).

A Vision Transformer (ViT) é uma simples adaptação do transformadores para tarefas de visão computacional, como classificação de imagens, em que a imagem de entrada é dividida em pedaços não sobrepostos. Estes alimentam um transformer convencional após uma camada linear Touvron et al. (2022). Mesmo que tais arquiteturas tenham ultrapassado ConvNets como modelos de classificação estado da arte, os ViTs convencionais enfrentam dificuldades quando aplicado a tarefas gerais de visão computacional como maior custo computacional. Neste cenário, um novo tipo de rede de convolução, a ConvNeXt, introduzida por Liu et al. (2022) surgiu para unir o poderio de uma rede de convolução convencional como a ResNeXt com os ViT, por meio dos Swin Transformers. O Swin Transformer aplica conceitos em que utiliza técnicas para melhorar o desempenho de um transformer convencional utilizando uma técnica de janelas deslocadas, tal técnica também se assemelha ao funcionamento de uma rede convolucional baseada em filtros, o que também trouxe sinergia na aplicação de tal tecnologia em uma rede de convolução convencional como a ResNeXt modificando-a de maneira a se espelhar no Swim Transformers o que levou ao surgimento de uma nova rede, batizada de ConvNeXt.

A ConvNeXt representou um avanço em direção ao estado da arte, na medida em que Demonstrou que pode alcançar o mesmo nível de escalabilidade que Transformadores de visão hierárquica enquanto manteve um design muito mais simples.

Neste contexto, este estudo tem como objetivo a análise dos modelos EfficientNet e ConvNext no desempenho da detecção de patologias em imagens de fundo de olho, visando efetuar um comparativo, demonstrando qualidades e desvantagens nas duas abordagens.

Este trabalho também visa, a contribuição no diagnóstico de doenças, efetuando diversas técnicas de otimização do aprendizado de redes neurais de convolução que realizarão a detecção de imagens de fundo de olho acometidas com patologias, viabilizando a futura utilização de tais modelos em campos médicos, auxiliando o dia a dia de profissionais e agilizando processos de triagem, contribuindo assim, com o diagnóstico precoce.

1.1 Trabalhos Relacionados

Em diversos estudos podemos perceber que modelos deep learning podem ser fundamentais para o auxílio de desafios médicos prevalentes. Em Wang et al. (2020), os autores propõem a construção de um modelo de rede neural convolucional com o propósito de classificar patologias em imagens de fundo de olho. A ODIR é utilizada como base de dados deste estudo, sendo comum imagens de fundo de olho de pacientes possuírem mais de uma patologia, caracterizando um problema multi-rótulo. Os autores propuseram uma transformação na base para que o problema fosse transformado em apenas 2 classes. Este modelo proposto utilizou diversas técnicas de ensemble e extrações de características e analisou diversas arquiteturas de redes neurais, sendo alocadas para propósitos de classificação. Dentre elas, a EfficientNet se destacou tanto pelo seu baixo custo computacional, quanto por ter alcançado os melhores resultados finais tendo uma score final de 0,7 (Média entre as métricas utilizadas no artigo: auc, kappa e F1 Score).

No trabalho de Bhawarkar, Yash et al. (2022) uma rede EfficientNet do tipo B5 foi utilizada para classificar retinopatia diabética em imagens de fundo de olho. Com foco na análise da severidade da doença, os autores treinaram a arquitetura escolhida para identificar 5 diferentes classes que representam a severidade da doença encontrada na imagem. A EfficientNet superou outras redes convolucionais de trabalhos relacionados ao estudo, obtendo uma acurácia de 0,80.

Khalil et al. (2019) objetivou a detecção e identificação do grau de severidade da Retinopatia Diabética, porém com o enfoque maior na busca de uma rede EfficientNet ideal por meio da otimização e data augmentation. Gao et al. (2022) demonstram como uma rede baseada em transformers pode ser escalável e utilizada de maneira análoga a uma CNN, construindo modelos híbridos capazes de unir o poderio de um ViT com o de redes de convoluções convencionais, demonstrando isso com o sólido desempenho em um problema como o de segmentação em que comumente é desencorajada a utilização de transformers.

A ConvNeXt também destaca-se na classificação de imagens médicas, sendo utilizada de forma semelhante á uma rede de convolução como a EfficientNet. Em Hassanien et al. (2022) o autor emprega uma ConvNeXt na identificação de tumores benignos ou malignos em imagens de ultrassom mamárias, mais especificamente na extração de

características de imagens de pacientes. Com ajuda de mecanismos de pooling e otimização de parâmetros, a rede proposta obteve uma acurácia de 91.66%, superando a própria EfficientNet e a MobileNetV3 que era a arquitetura que até então obtinha os melhores resultados para o mesmo tipo de problema com 87,42% de acurácia.

Finalmente, Li et al. (2022) nos introduz á uma poderosa rede de convolução com backbone alterado para empregar uma ConvNeXt. Tal arquitetura construída com o objetivo de segmentar e classificar núcleos sanguíneos de imagens microscópicas do sangue coletado de pacientes. Essa rede foi capaz de superar outras arquiteturas construídas com modelos estado da arte baseados em redes de convolução em um desafio chamado "CoNIC challenge 2022".

1.2 Objetivos

Neste contexto, este estudo tem como objetivo a análise e comparação de duas abordagens em um problema de classificação: modelos de redes neurais convolucionais como a EfficientNet e modelos híbridos que utilizam a tecnologia dos ViT por meio das ConvNeXt.

Este trabalho também visa, a contribuição na área médica fornecendo modelos capazes de efetuar a detecção entre imagens de fundo de olho patológicas ou saudáveis, objetivando acelerar a identificação de pacientes acometidos com doenças relacionadas á visão, contribuindo com diagnósticos precoces.

1.2.1 Objetivos Específicos

Destaca-se como objetivos específicos deste trabalho:

- Análise e comparação entre ConvNeXt e EfficientNet.
- Treinamento de redes capazes de detectar e identificar se uma imagem de fundo de olho está acometida com uma doença ou não.
- Utilização e análise de técnicas de data augmentation e de otimização de parâmetros.
- Estudo de técnicas de Fine Tuning.

2 Fundamentação Teórica

Este capítulo apresenta os conceitos explorados para o desenvolvimento do estudo como Deep Learning, redes neurais convolucionais, Vision Transformers, Swim Transformers bem como apresentar o funcionamento das redes de convolução estudadas: EfficientNet e ConvNeXt, além de contextualizar conceitos médicos que são fundamentais para o trabalho.

2.1 Imagens de fundo de olho

Com o envelhecimento da população, doenças como o glaucoma e diabetes naturalmente se tornarão mais comuns e evidentes com o passar dos anos.

Utilizando imagens de fundo de olho (Figura 2), sistemas de deep learning aparentaram serem capazes de diagnosticar glaucoma e outras doenças com maior acurácia que seres humanos. Auxiliando profissionais na classificação de imagens de fundo de olho, consequentemente reduzirá de forma drástica o custo, esforço e tempo necessário para detecção de tais patologias. Estudos na China e India indicam que a utilização de tais técnicas possuem um ótimo custo-benefício S.b c; Stalmans Ingeborgd; Ahmed Iqbal Ike K.e; Sng (2020).

Figura 2 – Imagem de fundo de olho de paciente com glaucoma.

2.2 Aprendizagem Profunda

Um dos maiores desafios de inteligências artificiais projetadas para aplicações do mundo real é a influência da variação dos mais diversos fatores que podem diferenciar a interpretação de cada dado a ser observado, como exemplo disso, temos imagens noturnas que a depender da iluminação, podem ser tomadas por pixels pretos ou, aproximando do estudo deste trabalho, uma imagem de fundo de olho muito lesionada, obstruindo a visão de estruturas como o disco óptico, tais fatores provocam ruídos e elevam o nível de abstração, dificultando a extração de características.

Deep Learning surge então com o poderio de solucionar tais problemas de abstração, por conta da introdução de representações de imagens que são expressadas a partir de representações mais simples das mesmas, ou seja, podemos identificar pessoas em imagens à partir do contornos de seus corpos, assim como podemos detectar patologias de fundo de olho à partir de seus padrões formados. Goodfellow, Bengio e Courville (2016).

No caso de classificações de imagens, é possível utilizar rótulos para identificar cada uma das diferentes classes de imagens de determinado dataset, que podem ser utilizados para organizar os dados de entrada de forma homogênea e medir o aprendizado do modelo de rede neural utilizado.

2.3 Rede Neural Convolucional (CNN)

Modelos de Deep Learning são implementados utilizando os conceitos de redes neurais artificiais, afim de simular o funcionamento de sinapses de um cérebro humano, essa tecnologia foi desenvolvida com o objetivo de reconhecer padrões, aprender das mais diversas características e atender aos mais diversos propósitos.

Uma rede neural convolucional se diferencia de tais modelos tradicionalmente usados em Deep learning, por conta de seu conceito de camadas (layers) e filtros, o que a torna uma ferramenta poderosa no manipulamento de imagens.

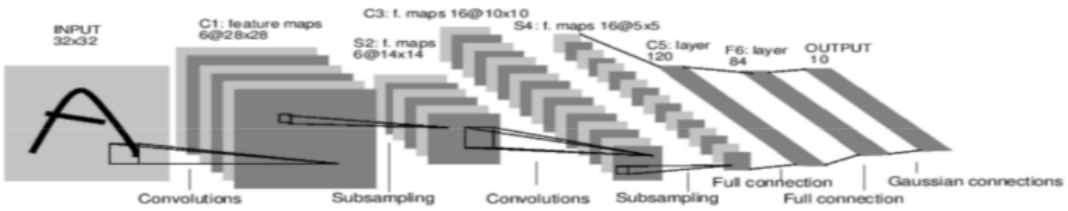


Figura 3 – Exemplo de CNN.

Uma CNN aplicada ao processamento de imagens funciona de maneira à receber imagens de forma matricial, a mesma imagem é passada por uma série de filtros, as

chamadas convoluções (Figura 4), cada uma delas realizada sobre a imagem pode resultar em diferentes transformações.

Os dados são tratados de forma espacial, considerando imagens como matrizes de pixels, filtros são aplicados à cada uma das matrizes, gerando como resultado os chamados mapas de características.

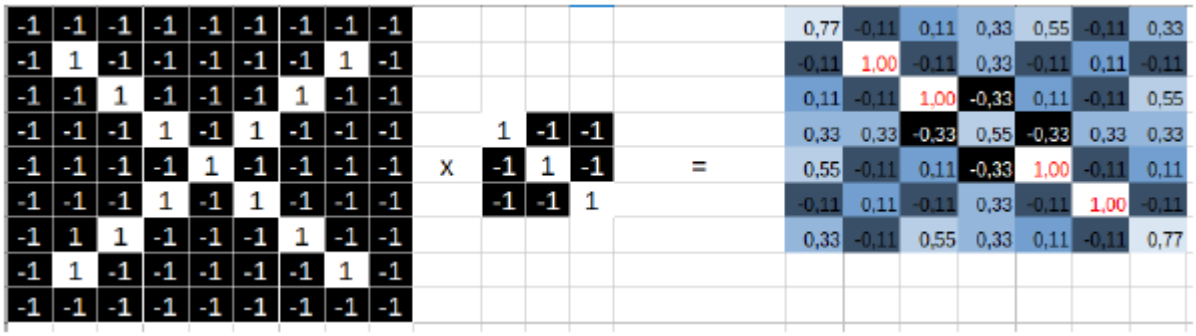


Figura 4 – Ilustração de convolução de um filtro sobre a matriz de pixels de uma imagem.

Além dos filtros convencionais, redes neurais de convolução possuem em sua arquitetura camadas batizadas de pooling (ou max pooling) (Figura 5), que reduzem o mapa de características criado, otimizando o processamento por conta da consequente redução de parâmetros efetuada pelas camadas pooling.

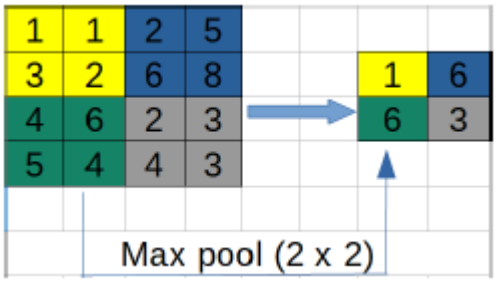


Figura 5 – Exemplo da aplicação de um filtro pooling sobre a matriz de pixels de uma imagem.

Após todo o processo de filtragem da imagem pelas diferentes camadas, características são analisadas de forma sólida pelas chamadas funções de ativação, que são aplicadas ao valor resultante das convoluções, tornando a rede, além de outras utilidades, capaz de identificar ou recriar a imagem ou reconhecer as probabilidades de pertencimento da imagem à certa classe.

2.3.1 EfficientNet

A EfficientNet é um modelo de rede neural proposta por (TAN; LE, 2019) que implementa um método de escalabilidade composta, segundo o estudo, mostrou-se ser de critica importância equilibrar largura, profundidade e resolução da rede na busca pela

escalabilidade de parâmetros e tal equilíbrio poderia ser alcançado dimensionando cada um deles com razão constante.

O método de escalabilidade composta, busca implementar tal equilíbrio, de forma a, caso necessário, utilizar-se 2^N custos computacionais a mais, a rede poderia acrescentar α^N de largura, β^N de profundidade e ϵ^N de resolução Tan e Le (2019). A (Figura 6) exemplifica diferentes métodos de escalabilidades em comparação com o proposto pelo trabalho da EfficientNet.

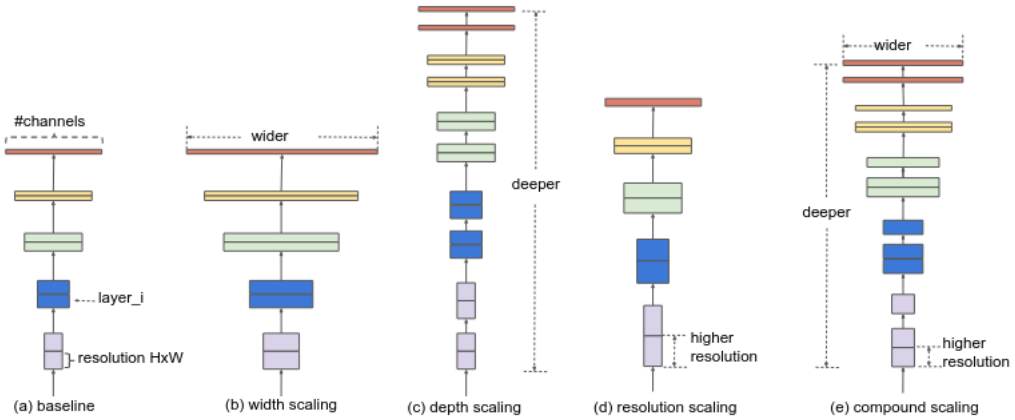


Figura 6 – Comparações entre métodos de escalabilidade e escalabilidade composta.

Naturalmente, maiores resoluções de imagem entregam maior visibilidade das mesmas, tornando mais fácil a extração de características e facilitando o aprendizado. Observou-se que para maiores resoluções, uma rede neural deve ser mais profunda de modo que campos receptivos possam capturar uma maior quantidade de pixeis e mais larga para capturar padrões mais refinados com a quantidade maior de em imagens de alta resolução.

A abordagem proposta de escalabilidade composta, propõe aumentar cada uma das dimensões da rede (profundidade, largura e resolução) de modo uniforme (Figura 7).

$$\begin{aligned} \text{depth: } & d = \alpha^\phi \\ \text{width: } & w = \beta^\phi \\ \text{resolution: } & r = \gamma^\phi \\ \text{s.t. } & \alpha \cdot \beta^2 \cdot \gamma^2 \approx 2 \\ & \alpha \geq 1, \beta \geq 1, \gamma \geq 1 \end{aligned}$$

Figura 7 – Método de escalabilidade composta, onde θ é o coeficiente para escalar uniformemente cada uma das 3 dimensões sendo, α, β e ϵ constantes determinadas por uma busca de parâmetros utilizando *grid search*.

A proposta de escalabilidade da EfficientNet foi avaliada utilizando redes neurais convolucionais já existentes, buscando maximizar a acurácia para quaisquer restrições de recursos, como representado pela fórmula da (Figura 8) e buscando também otimizar

custos computacionais (FLOPS). A EfficientNet B0 (Figura 9) então surge após a busca na escalabilidade dos parâmetros de uma rede baseada na já existente MobileNet.

$$
\begin{aligned}
\max_{d,w,r} \quad & Accuracy(\mathcal{N}(d,w,r)) \\
s.t. \quad & \mathcal{N}(d,w,r) = \bigodot_{i=1\ldots s} \hat{\mathcal{F}}_i^{d\cdot\hat{L}_i}\left(X_{\langle r\cdot\hat{H}_i, r\cdot\hat{W}_i, w\cdot\hat{C}_i\rangle}\right) \\
& \text{Memory}(\mathcal{N}) \leq \text{target\_memory} \\
& \text{FLOPS}(\mathcal{N}) \leq \text{target\_flops}
\end{aligned}
$$

Figura 8 – Fórmula que representa o problema de maximização da acurácia para quaisquer restrições de recursos, onde w, d e r são os coeficientes para escalar, respectivamente, largura, profundidade e resolução e $\hat{\text{F}}i$, $\hat{\text{L}}i$, $\hat{\text{H}}i$, $\hat{\text{W}}i$ e $\hat{\text{C}}i$ parâmetros pré-definidos pelo modelo base.

Stage i	Operator $\hat{\mathcal{F}}_i$	Resolution $\hat{H}_i \times \hat{W}_i$	#Channels $\hat{C}_i$	#Layers $\hat{L}_i$
1	Conv3x3	224×224	32	1
2	MBConv1, k3x3	112×112	16	1
3	MBConv6, k3x3	112×112	24	2
4	MBConv6, k5x5	56×56	40	2
5	MBConv6, k3x3	28×28	80	3
6	MBConv6, k5x5	14×14	112	3
7	MBConv6, k5x5	14×14	192	4
8	MBConv6, k3x3	7×7	320	1
9	Conv1x1 & Pooling & FC	7×7	1280	1

Figura 9 – EfficientNet B0. Cada linha da tabela descreve um estágio de convolução i com $\hat{\text{L}}i$ camadas, com resolução de input $\langle\hat{\text{H}}i, \hat{\text{W}}i\rangle$ e canais de output $\hat{\text{C}}i$.

2.3.2 ConvNeXt

A ConvNeXt (LIU et al., 2022) é uma arquitetura de rede neural que foi desenvolvida para lidar com tarefas de visão computacional, como classificação de imagens e detecção de objetos. A principal diferença entre a ConvNeXt e outras arquiteturas CNN é que ela utiliza uma combinação de camadas de convolução e camadas que se espelham no funcionamento de transformadores (ViTs), mais especificamente, Swim Transformers, o que permite que ela aprenda representações mais abstratas e genéricas a partir de dados de imagem.

Os vision transformers são arquiteturas baseadas em transformadores, um tipo de rede neural que foi originalmente desenvolvida para processamento de linguagem natural. Transformadores são uma classe de redes neurais que são baseadas em uma arquitetura de "attention", ou seja, elas usam mecanismos de atenção para focar em partes específicas dos dados de entrada durante o processamento. Isso permite que os transformadores aprendam representações mais abstratas e genéricas a partir dos dados de entrada.

Em um ViT, as camadas de transformadores são usadas para processar imagens de entrada e aprender representações de características de imagem de maneira eficiente. As imagens são primeiro divididas em pedaços menores (Figuras 10 e 11) e cada pedaço é processado por uma camada de transformadores (Figura 12). As camadas de transformadores aprendem a combinar esses pedaços em uma representação mais abstrata da imagem. Esse processo é repetido várias vezes, permitindo que o modelo aprenda representações cada vez mais abstratas da imagem.

Figura 10 – Transformação da imagem em vários pedaços dela mesma.

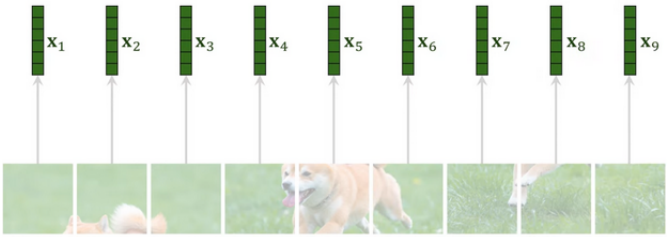


Figura 11 – Processo de vetorização, em que cada pedaço se transforma em um vetor de características para facilitar a representação dos mesmos para a rede.

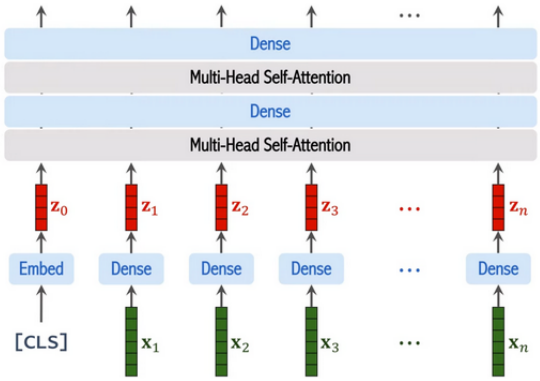


Figura 12 – Camadas de attention e de ativação de um ViT que irão aprender características específicas de cada pedaço da imagem (vetores).

O Swin Transformer (SwimT) é um modelo de aprendizado de máquina proposto por (LIU et al., 2021) que visa melhorar a performance de modelos de transformadores visuais hierárquicos. O SwimT também se aproxima do funcionamento de uma rede de convolução baseada em filtros devido a sua técnica de processar imagens.

A arquitetura SwimT utiliza uma técnica chamada janelas deslocadas para dividir a imagem de entrada em pedaços menores (Figura 13), o que permite que o modelo processe cada parte da imagem de forma independente. Isso é útil porque permite que o modelo aprenda características mais específicas em cada parte da imagem, o que pode levar a uma melhor compreensão global da imagem.

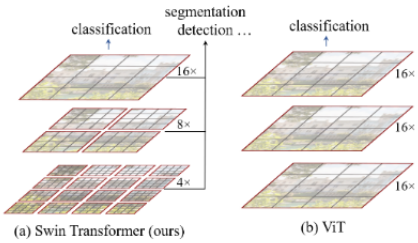


Figura 13 – Representação do funcionamento da técnica de janelas deslocadas em comparação com o funcionamento de um ViT que mantém o processamento da imagem de maneira mais global que um SwimT.

À partir da técnica de janelas deslocadas utilizadas, o Swim Transformer pôde resolver um problema de escalabilidade de ViTs comuns (Figura 14), além disso, o uso de janelas deslocadas também permite que o modelo tenha uma visão mais ampla da imagem, o que pode ser útil para tarefas de reconhecimento de objetos onde é importante ter uma compreensão do contexto em que os objetos estão inseridos.

$$\Omega(\textbf{MSA}) = 4hwC^2 + 2(hw)^2C, \qquad (1)$$

$$\Omega(\textbf{W-MSA}) = 4hwC^2 + 2M^2hwC, \qquad (2)$$

Figura 14 – Comparação entre a fórmula de escalabilidade de custo computacional de um bloco *Muti-head self attention* (ViT) (1) e um bloco *Windowned Muti-head self attention* (SwimT) (2). Perceba que em relação à altura (h) e largura (w) a escalabilidade passou de quadrática (1) para linear (2) em relação à essas duas variáveis de resolução da imagem.

Finalmente, inspirado pelas recentes modernizações de redes de convolução e a ascensão dos ViT na resolução de tarefas de visão computacional, a ConvNext foi proposta por Liu et al. (2022), buscando unir o poderio de redes de convolução estado da arte com a tecnologia dos transformers, criando uma mistura entre ResNext (XIE et al., 2016) e Swim Transformers.

Em um primeiro momento, os autores propuseram o que chamaram de "ResNext-fy", onde uma ResNet-50 passaria por alterações em que cada um de seus filtros convolucionais passariam a operar separados em diferentes grupos de convolução, aumentando assim, a largura da rede (Figura 15). Seguindo a estratégia proposta da ResNeXt em Xie et al. (2016), também foram aumentados os números de canais de 64 para 96. Com estas alterações, a rede se aproximou do funcionamento de uma rede self-attention e de ViTs convencionais.

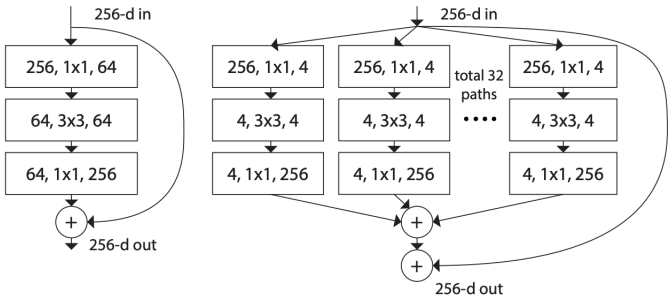


Figura 15 – Na esquerda, um bloco da ResNet-50, á direita, um bloco da ResNeXt com cardinalidade = 32, onde mantém a mesma complexidade.

Um design que não poderia ser esquecido na construção da ConvNeXt é o gargalo invertido (Figura 16), que é muito comum em ViTs, onde a hidden dimension do bloco MLP é 4 vezes maior que as dimensões de input. Este também é uma estrutura similar à utilizadas em convnets como a MobileNetV2, os kernels também foram aumentados de 3x3 para 7x7, significativamente maiores aos da ResNeXt, buscando aproximação com Swim Transformers e melhorando a performance da rede.

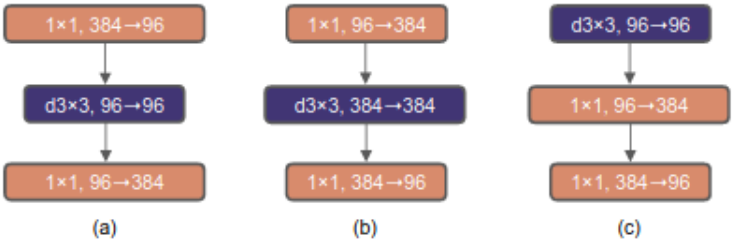


Figura 16 – Gargalo invertido onde em (a) temos um gargalo da ResNeXt, em (b) a camada hidden (ao centro) teve suas dimensões aumentadas em 4 vezes e em (c) a camada hidden teve sua posição avançada para se adaptar às futuras modificações de kernel.

Em uma visão micro, também foram alteradas as funções de ativação de ReLU para GELU, e organizando uma ativação para cada bloco, utilizando apenas um bloco de normalização em vez de três e substituindo Batch normalizations (BN) por Layer normalizations (LN) (Figura 17).

Finalmente, a arquitetura ConvNext está sumarizada pela (Figura 18).

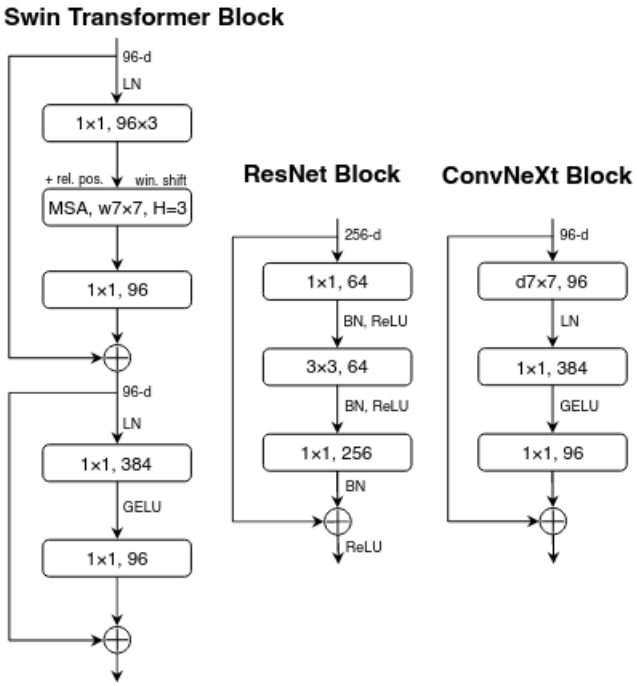


Figura 17 – Comparação entre um bloco de um Swin Transformer com um da ResNet com um da ConvNeXt após todas as modificações efetuadas na ResNeXt.

	output size	• ResNet-50	• ConvNeXt-T	○ Swin-T
stem	56×56	7×7, 64, stride 2 3×3 max pool, stride 2	4×4, 96, stride 4	4×4, 96, stride 4
res2	56×56	[1×1, 64 3×3, 64 1×1, 256] × 3	[d7×7, 96 1×1, 384 1×1, 96] × 3	[1×1, 96×3 MSA, w7×7, H=3, rel. pos. 1×1, 96 1×1, 384 1×1, 96] × 2
res3	28×28	[1×1, 128 3×3, 128 1×1, 512] × 4	[d7×7, 192 1×1, 768 1×1, 192] × 3	[1×1, 192×3 MSA, w7×7, H=6, rel. pos. 1×1, 192 1×1, 768 1×1, 192] × 2
res4	14×14	[1×1, 256 3×3, 256 1×1, 1024] × 6	[d7×7, 384 1×1, 1536 1×1, 384] × 9	[1×1, 384×3 MSA, w7×7, H=12, rel. pos. 1×1, 384 1×1, 1536 1×1, 384] × 6
res5	7×7	[1×1, 512 3×3, 512 1×1, 2048] × 3	[d7×7, 768 1×1, 3072 1×1, 768] × 3	[1×1, 768×3 MSA, w7×7, H=24, rel. pos. 1×1, 768 1×1, 3072 1×1, 768] × 2
FLOPs		4.1×10^9	4.5×10^9	4.5×10^9
# params.		25.6×10^6	28.6×10^6	28.3×10^6

Figura 18 – Comparação entre a arquitetura final ConvNext com a ResNet-50 e a de Swim Transformer.

3 Metodologia

Este capítulo apresenta a metodologia proposta para a realização das avaliações do desempenho do estudo, buscando igualar condições de input de dados e treinamento e otimizar o melhor possível a base de dados utilizada por este estudo, destacando também as técnicas de treinamento e augmentation de dados testadas neste trabalho.

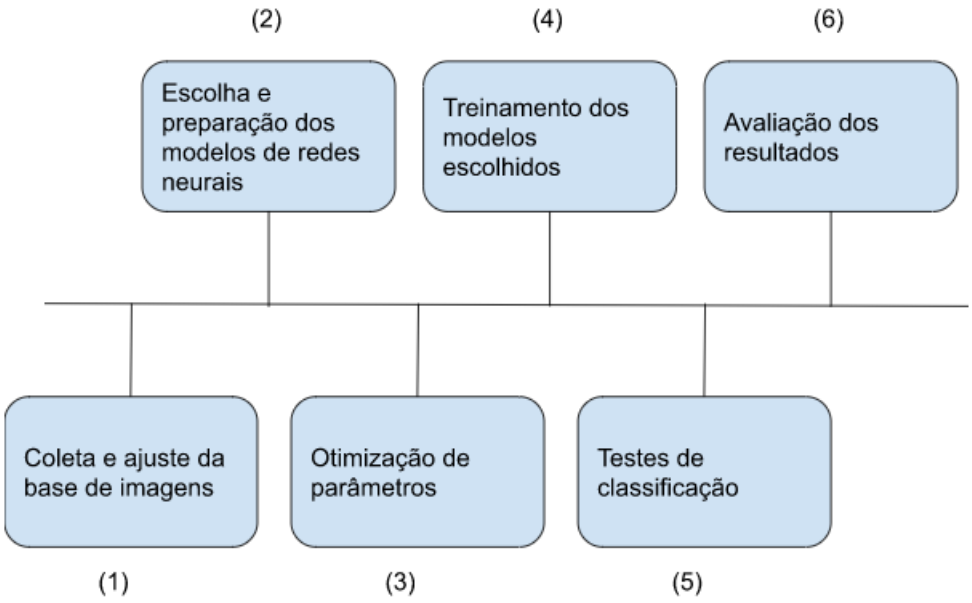


Figura 19 – Fluxo da metodologia proposta.

3.1 Coleta de dados

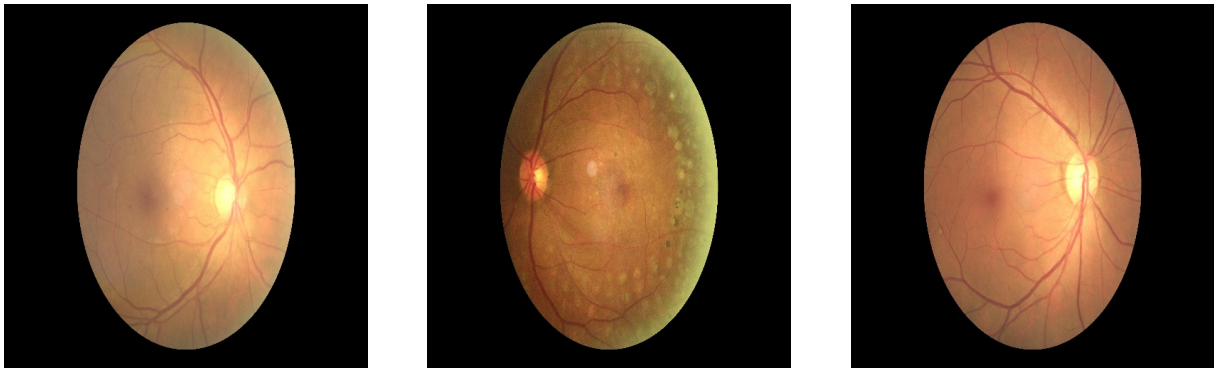
O dataset *Ocular Disease Recognition* (ODIR-5K) UniversityIAI-PKU (2019) foi escolhida para avaliação desse estudo. É um banco de dados oftalmológico estruturado de 5000 pacientes (Figura 20) com idade, fotografias coloridas de fundo de olho dos olhos esquerdo e direito e palavras-chave de diagnóstico dos médicos, totalizando 8000 imagens de fundo de olho, entre elas, 6392 rotuladas com suas respectivas patologias com as quais o paciente está acometido ou com identificação de fundo de olho normal para pacientes saudáveis (Tabela 1).

3.2 Ajuste dos dados

Na análise inicial dos dados, notou-se que a base Odir é composta em sua grande maioria por imagens Multi-Label, onde muitos dos pacientes possuem mais de uma patologia (Figura 21). Caso fossem efetuados tratamentos para uma abordagem de classificação

Tabela 1 – ODIR - Rótulos de imagens

Rótulos	**Número de imagens**
Fundo de olho normal	2874
Retinopatia	1489
Glaucoma	218
Catarata	262
Degeneração Macular	289
Hipertensão	206
Miopia	227
Multi rótulos	827



Retinopatia diabética Glaucoma Normal

Figura 20 – Exemplos de imagens da base Odir com seus respectivos rótulos.

Multi-label, o input de dados para treinamento resultaria em um universo menor do que o potencial a ser explorado no universo da base, o que tornaria desafiadora e porventura improdutiva a análise do desempenho das arquiteturas.

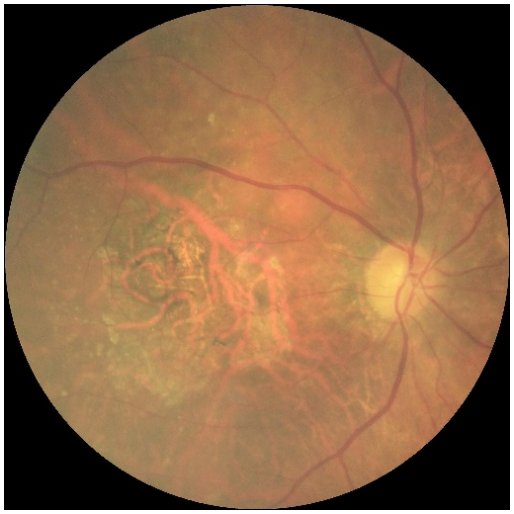


Figura 21 – Imagem de fundo de olho de paciente acometido com glaucoma e degeneração da mácula.

Logo, a classificação binária foi adotada por este estudo, percebeu-se que tal abordagem otimizaria melhor os dados, de forma a dividir as classes em apenas duas:

pacientes com fundo de olho sem patologias e pacientes com fundo de olho com patologias (Tabela 2), permitindo a utilização de todo o universo de amostragem fornecido pelo dataset e se aproximando de uma implementação mais funcional, pois uma inteligência artificial programada para detectar com precisão se um paciente está acometido com doenças relacionadas ao fundo de olho, reduziria drasticamente o tempo e custo necessários para efetuar diagnósticos médicos.

Tabela 2 – Organização de imagens efetuada por este estudo

Classe	**Número de imagens**
Fundo de olho normal	2874
Fundo de olho patológico	3518

3.2.1 Balanceamento de classes de imagens

a Em um processo de aprendizado de uma rede neural, é essencial que os dados sejam ajustados de maneira à torná-lo o mais proveitoso possível. Em um problema de classificação, o balanceamento de dados entre as classes é um ajuste interessante à ser feito, pois um número igual de imagens de cada tipo evita o enviesamento e aumenta o alcance de aprendizado da rede, evitando assim o overfitting.

Primeiramente, as 2 classes ilustradas pela Tabela 2, foram balanceadas com a retirada de 644 imagens de fundo de olho doentes, resultando em 2874 imagens para cada classe. O segundo passo efetuado, foi a divisão entre treino, validação e teste, em que 60% foram separadas para treino, das 40% que restaram, 80% foram separadas para validação e 20% para teste, mantendo o balanceamento de classes entre cada uma das divisões (Tabela 3).

Tabela 3 – Divisão treino, teste e validação

Divisão	**Número de imagens**
Treino	3448
Validação	1840
Teste	460

3.2.2 Data Augmentation

Para preparar a base de dados para o treinamento em si, as imagens que foram selecionadas para o treino passaram por etapas de data augmentation, afim de aumentar a quantidade de imagens e a variedade das mesmas.

Utilizando a biblioteca Albumentations (BUSLAEV A. PARINOV; KALININ, 2018), cada uma das imagens da divisão de treino passaram por um pipeline em que foram alteradas de formas randômicas. Sendo as possibilidades de alterações: inversão horizontal,

inversão vertical e dupla alteração de contraste, como demonstradas pelas (Figuras 22 e 23).

Figura 22 – Exemplo de imagem invertida horizontalmente.

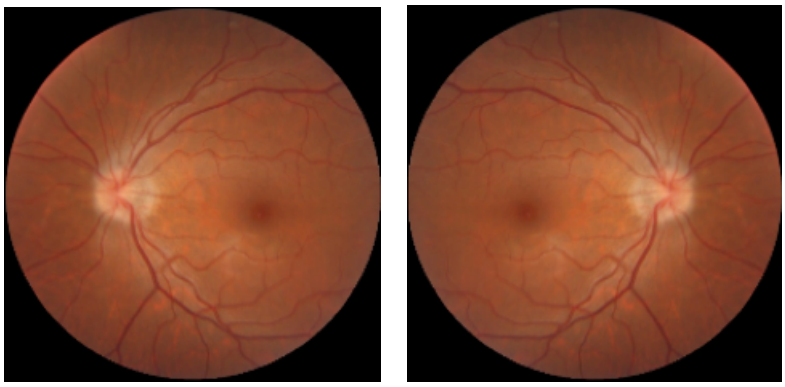


Imagem original Após augmentation.

Figura 23 – Exemplo de imagem invertida horizontalmente, verticalmente e que sofreu alterações de contraste.

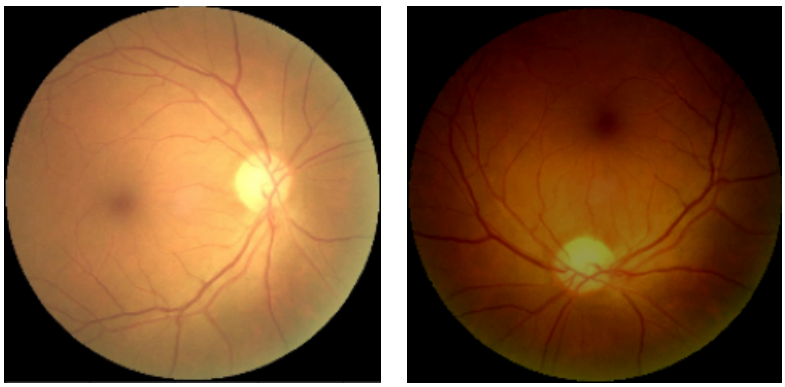


Imagem original Após augmentation.

À partir do data augmentation, foram criadas 3448 novas imagens, sendo metade delas da classe de fundo de olho normal e a outra metade da classe de fundo de olho patológico, introduzidas para enriquecer o dataset de treino dividido anteriormente. A confiabilidade dos dados foram asseguradas com a randomização de modificações e com as imagens passando pelo pipeline de modificação apenas uma vez cada, as imagens também foram redimensionadas para 224x224 para se adequar à limitações técnicas de consumo de RAM no treinamento e por ser um input médio ideal entre ConvNeXt e EfficientNet. O número de imagens do dataset final está exposto na (Tabela 4).

Tabela 4 – Divisão final de treino, teste e validação.

Divisão	**Número de imagens**
Treino	6896
Validação	1840
Teste	460

3.3 Escolha e preparação dos modelos de redes neurais

Como citado anteriormente, a EfficientNet (TAN; LE, 2019) e a Convnext (LIU et al., 2022) foram as arquiteturas utilizadas por este estudo com o objetivo de analisar e comparar modelos estado da arte como estes em um problema de classificação de patologia em imagens de fundo de olho.

Foram utilizadas as implementações dos respectivos modelos da biblioteca keras, disponível para uso com TensorFlow, cada uma de suas versões e possibilidades de parâmetros foram avaliadas na etapa de otimização.

É importante destacar que a preparação de dados foi feita da mesma maneira em cada um dos treinamentos dos dois modelos e otimizações, objetificando igualar os cenários de treinamento e teste de cada uma das arquiteturas além de buscar uma análise do desempenho de ambas de forma fidedigna e livre de vieses na entrada de dados.

Afim de tirar proveito de transfer learning, foram utilizados modelos pré-treinados na base Imagenet (DENG et al., 2009), em que o topo de cada um dos modelos foi substituído por camadas Dense de ativação em que a função de ativação utilizada foi a Relu, camadas de Dropout e uma camada de classificação. Naturalmente, nas etapas de treinamentos e testes utilizando Fine Tuning, os modelos foram preparados de maneira à efetuar tais técnicas.

No Deep Fine Tuning, os pesos dos modelos originais foram congelados totalmente em um primeiro momento do treinamento em que foram treinadas apenas as camadas de ativação e classificação e no caso do Shallow Fine Tuning, apenas parte das camadas dos modelos originais foram congeladas na primeira etapa do treinamento em que foram treinadas as camadas do topo do modelo.

3.4 Otimização de parâmetros

No processo de treinamento de uma rede neural é importante a análise de cada uma de suas combinações de parâmetros, buscando otimizar o aprendizado o máximo possível, melhorando assim, os resultados finais na etapa de testes. Neste estudo, a biblioteca Keras Tuner (O'MALLEY et al., 2019) foi utilizada na implementação do algorítimo de busca utilizado, o Hyperband (LI et al., 2018), algoritmo adaptativo de busca randômica de parâmetros ideais.

Foram testadas as mais diversas combinações de parâmetros envolvendo em especial, os diversos tipos de modelos EfficientNet e ConvNeXt disponíveis (Tabela 5), tais arquiteturas se diferem essencialmente na quantidade de parâmetros e no tipo de base Imagenet em que os mesmos foram treinados.

No caso de backbones Base e Large da ConvNeXt, existem dois tipos de cada, sendo um deles treinado na Imagenet 21k (21000 imagens) e passado por Fine Tuning posterior na Imagenet 1k (1000 imagens) e o segundo apenas treinado na Imagenet 1k, tais backbones foram caracterizados com o número de imagens das bases Imagenet em que foram treinados em seus nomes.

Estas arquiteturas e outras da (Tabela 5) foram testadas na etapa de otimização de parâmetros, afim de buscar o backbone ideal para a base de imagens e o problema de classificação do estudo. Os backbones da EfficientNet diferem-se basicamente na quantidade de parâmetros, os mesmos também possuindo seus pesos pré-treinados utilizando a base de dados da Imagenet.

Tabela 5 – Backbones ConvNeXt e EfficientNet testadas na otimização.

ConvNeXt
Tiny
Small
Base 1k
Base 21k 1k
Large 1k
Large 21k 1k
xLarge 21k 1k

EfficientNet
EfficientNetB0
EfficientNetB1
EfficientNetB2
EfficientNetB3
EfficientNetB4
EfficientNetB5
EfficientNetB6
EfficientNetB7

Os outros parâmetros testados estão expostos na (Tabela 6).

Tabela 6 – Demais combinações de parâmetros testadas. Incluindo o número de camadas de ativação adicionada, o dropout entre elas e o número de neurônios em cada camada acrescentada ao modelo original.

Otimizadores	**Taxas de aprendizado**	**Camadas**	**Dropout**	**Nº de neurônios**
Sgd	10^{-2}	1	0.2	16
Rmsprop	10^{-3}	2	0.5	32
Adam	10^{-4}	3	0.8	64
	10^{-5}			128
	10^{-6}			256
	10^{-7}			

Como resultado das etapas de otimização Hyperband efetuadas individualmente para modelos ConvNeXt e Efficient, obteve-se os seguintes modelos otimizados da (Tabela 7). Estes foram as configurações de modelos utilizadas nas etapas de testes e avaliações finais.

3.5 Treinamento dos modelos escolhidos

Após configurados os modelos com combinações de parâmetros ideais encontradas na etapa de otimização, preparou-se ferramentas para auxiliar o treinamento, estas efetuam

Tabela 7 – Classificação de imagens de fundo de olho patológicas - Modelos otimizados

Parâmetro	**ConvNeXt**	**EfficientNet**
Backbone	xLarge 21k 1k	B2
Otimizador	Adam	Adam
Taxa de aprendizado	10^{-6}	10^{-5}
Nº de camadas	2	1
Dropout	0.8	0.8
Nº Neurônios por camada	16	256

técnicas que otimizam a utilização dos dados utilizados, como o balanceamento de batches e o data augmentation. Os treinamentos foram efetuados em 15 épocas cada com um tamanho de batch de 8 para cada etapa.

A cada época de um treinamento de rede neural, um pedaço (batch) da base de dados é escolhido para que o modelo aprenda características sobre cada uma das imagens do batch, para garantir o balanceamento entre as classes em cada época do treinamento e com isso, assegurar a análise das arquiteturas estudadas, a biblioteca Imbalanced-learn (LEMAîTRE; NOGUEIRA; ARIDAS, 2017) foi utilizada para criar um data generator que balanceia batches à cada época.

Naturalmente, o treinamento durante a avaliação de técnicas de Fine Tuning foi adaptado para duas etapas. No Deep Fine Tuning o modelo encontrado na etapa de otimização foi congelado por inteiro, com exceção das camadas de ativação e classificação enquanto no Shallow Fine Tuning os pesos das camadas do topo do modelo original também permaneceram descongeladas.

Em um primeiro momento, apenas as camadas descongeladas foram treinadas, após este treinamento, descongelou-se todas os outros pesos das arquiteturas e as mesmas passaram por uma segunda etapa de treinamento para então passarem por uma avaliação final de resultados.

3.6 Avaliação de resultados

As métricas utilizadas para avaliação do aprendizado dos modelos de classificação foram *Precision*, *Recall* e *F1-score* (Figura 24).

Figura 24 – Métricas utilizadas na avaliação de resultados.

$$\text{Precision} = \frac{TP}{TP+FP} \qquad \text{Recall} = \frac{TP}{TP+FN} \qquad F1\ score = 2 * \frac{Precision\ * Recall}{Precision + Recall}$$

Tais métricas são aplicadas na fase de testes, onde as redes são expostas à divisão teste da tabela 4 e tentarão efetuar uma previsão da classe de cada imagem.

A métrica *Precision* busca encontrar qual proporção de identificações positivas (*TP* - verdadeiros positivos) a rede foi capaz de acertar em comparação com as positivas identificadas erroneamente (*FP* - falsos positivos), *Recall* verifica a proporção de identificações positivas em comparação com as negativas identificadas erroneamente (*FN* - falsos negativos) e *F1-Score* efetua uma avaliação matemática entre *Precision* e *Recall*, para obter uma métrica que avalie o resultado entre elas.

4 Resultados

Este capítulo apresenta as avaliações efetuadas de cada um dos testes efetuados, bem como discorre o impacto de cada arquitetura nos resultados obtidos e como cada uma delas se comportou com as diferentes técnicas utilizadas.

4.1 Etapa de treinamento das redes inteiras

Nesta etapa, cada um dos pesos das duas redes foram treinados com a base de dados apresentada anteriormente, afim de obter a capacidade de identificar entre imagens de fundo de olho patológicas ou saudáveis. O treinamento foi efetuado em 15 épocas à um batch size de 854 imagens por época e os modelos utilizados foram parametrizados de acordo com a etapa de otimização efetuada anteriormente.

A Tabela 8 mostra os resultados após a predição efetuada pelas redes descritas na etapa inicial.

Tabela 8 – Classificação de imagens de fundo de olho patológicas - Modelos otimizados

Modelo	Precision	Recall	F1-Score
ConvNeXt	0,64	0,63	0,62
EfficientNet	0,73	0,72	0,72

Já percebe-se pelas média entre *Precision*, *Recall* e *F1-Score* de 0,63 e 0,72, uma capacidade dos modelos treinados de aprender características do problema em questão e superar uma probabilidade de 50% na predição, nesta avaliação, a EfficientNet superou a ConvNeXt em termos de resultados.

Outro ponto interessante de relatar é o fato de que nesta etapa, a EfficientNet levou ligeiramente menos tempo para ser treinada (Tabela 9), entretanto o backbone ConvNeXt utilizado tinha uma quantidade de parâmetros consideravelmente maior o que em teoria, poderia levar a rede a ser mais lenta, o que não necessariamente foi o caso.

Tabela 9 – Tempo levado para treinar por época - Modelos otimizados

Modelo	Tempo médio por época (segundos)
ConvNeXt	330s
EfficientNet	320s

Nesta etapa incial já pôde-se perceber que a ConvNeXt consegue se aproximar de resultados de uma rede neural estado da arte. Também pode-se concluir que as técnicas de processamento da base de dados e de balanceamento de batches já puderam levar a

EfficientNet treinada à resultados próximos ou superiores à estudos semelhantes como o de (WANG et al., 2020) e (Bhawarkar, Yash et al., 2022).

4.2 Etapa de treinamento utilizando técnicas de Fine Tuning

Nesta etapa, as técnicas de Fine Tuning (Deep Fine Tuning e Shallow Fine Tuning) foram testadas e avaliadas. O treinamento do topo da rede foi efetuado em 5 épocas e após o descongelamento de pesos, os modelos passaram por mais uma etapa de treinamento realizado em 10 épocas cada um dos treinamentos foi efetuado com um tamanho de batch de 854 imagens por época e os modelos foram parametrizados de acordo com a etapa de otimização.

As Tabelas 10 e 11 apresentam os resultados após a predição efetuada pelas redes descritas nesta etapa.

Tabela 10 – Classificação de imagens de fundo de olho patológicas - Deep Fine Tuning

Modelo	Precision	Recall	F1-Score
ConvNeXt	0,68	0,66	0,66
EfficientNet	0,73	0,72	0,71

Tabela 11 – Classificação de imagens de fundo de olho patológicas - Shallow Fine Tuning

Modelo	Precision	Recall	F1-Score
ConvNeXt	0,71	0,71	0,71
EfficientNet	0,79	0,78	0,78

Pode-se perceber que as redes se comportaram bem com as técnicas de Fine Tuning analisadas, melhorando suas métricas e em termos de tempo de processamento mantendo os resultados da (Tabela 9).

A técnica que obteve os melhores resultados foi a Shallow Fine Tuning com uma média entre *Precision*, *Recall* e *F1-Score* de 0,71 para a ConvNeXt e 0,79 para a EfficientNet (Tabela 11), indo de acordo com a avaliação da etapa de treinamento dos modelos como um todo, a ConvNeXt também pôde se aproximar da EfficientNet quando uma técnica de Fine Tuning é aplicada.

No problema de classificação estudado, levando em conta os melhores resultados obtidos por este estudo, pode-se verificar que a EfficientNet treinada pôde se aproximar ou superar resultados de trabalhos semelhantes em que tal modelo também foi utilizado, como mostra a (Tabela 12) o que é um indício de que os métodos de pré-processamento de dados e de treinamento tiveram um impacto positivo no resultado final.

É importante destacar que os resultados de estudos relacionados servirem de base para este estudo, os mesmos foram obtidos em circunstâncias de treinamento, objetivos e configurações de rede diferentes da rede EfficientNet obtida por neste trabalho.

Tabela 12 – Comparação entre os resultados da EfficientNet obtida neste estudo com outros modelos EfficientNet obtidos por estudos realizados sobre problemas de classificação semelhantes.

Modelo obtido	(WANG et al., 2020)	(Bhawarkar, Yash et al., 2022)
0,79	0,70	0,80

5 Conclusão

Este trabalho pôde explorar a análise de redes neurais estado da arte em um problema de classificação de patologia em imagens de fundo de olho, tais modelos, ConvNeXt e EfficientNet que tinham duas diferentes abordagens baseadas, respectivamente, em Vision Transformers e Redes Neurais de Convolução escaláveis, também foi possível testar e verificar o impacto de diferentes métodos de processamento de imagens e de balanceamento na fase de treinamento que tiveram um impacto significativo nos resultados finais e por fim, analisar técnicas de treinamento como Fine Tuning e de otimização de parâmetros com Hyperband.

Analisando os resultados numéricos das predições efetuadas, conclui-se que os modelos utilizados neste estudo foram capazes de aprender características de imagens de fundo de olho, resultando em inteligências artificiais capazes de detectar na maioria das vezes se uma imagem de fundo de olho é patológica ou não.

Em especial, a EfficientNet obtida pelo estudo, pôde se aproximar ou superar em média de métricas estudos em cima de problemas de classificação semelhantes como os de (WANG et al., 2020) e (Bhawarkar, Yash et al., 2022). Este estudo também deixa como legado, resultados da ConvNeXt obtida, que se aproximou de resultados obtidos por uma rede estado da arte como a EfficientNet.

O comportamento dos modelos no que se refere à escalabilidade de resoluções de imagens é um ponto importante a ser analisado por trabalhos posteriores e que não pôde ser explorado por este trabalho por conta de limitações de hardware, sobretudo a ConvNeXt que, em teoria, pode obter resultados mais expressivos quanto maior for a resolução de imagens, devido à sua estrutura que mimifica um Swim Transformer que se beneficia de maior riqueza de detalhes em imagens para o aprendizado de características de forma mais profunda, o que pode ser obtido com maiores resoluções ao sacrifício de um maior custo computacional.

No trabalho de classificar imagens de fundo de olho patológicas, a EfficientNet superou a outra rede estudada em resultados numéricos, o que mostra o poderio de uma rede de convolução estado da arte convencional, baseada em técnicas de escalabilidade de parâmetros e que também deve melhorar seus resultados ao ser exposta à maiores resoluções. Entretanto não deve-se subestimar o potencial de uma rede híbrida como a ConvNeXt que também obteve resultados satisfatórios e ainda possui muito à ser analisada e certamente é uma das arquiteturas mais promissoras desta década.

De forma eficiente, este trabalho foi capaz de analisar e treinar dois modelos de redes neurais de convolução estado da arte capazes de identificar com satisfatória precisão

imagens de fundo de olho patológicas ou saudáveis, deixando como legado toda a análise de parâmetros, técnicas de treinamento e processamento de imagens que podem ser utilizadas ou exploradas por trabalhos futuros e por fim, abrindo espaço para a utilização de tais modelos na análise médica, facilitando o trabalho de profissionais da saúde e reduzindo custos e tempo de espera.

Referências

Bhawarkar, Yash; Bhure, Kaustubh; Chaudhary, Vinayak; Alte, Bhavana. Diabetic retinopathy detection from fundus images using multi-tasking model with efficientnet b5. *ITM Web Conf.*, v. 44, p. 03027, 2022. Disponível em: <https://doi.org/10.1051/itmconf/20224403027>. Citado 4 vezes nas páginas 15, 35, 36 e 37.

BUSLAEV A. PARINOV, E. K. V. I. I. A.; KALININ, A. A. Albumentations: fast and flexible image augmentations. *ArXiv e-prints*, 2018. Citado na página 28.

DENG, J.; DONG, W.; SOCHER, R.; LI, L.-J.; LI, K.; FEI-FEI, L. Imagenet: A large-scale hierarchical image database. In: IEEE. *2009 IEEE conference on computer vision and pattern recognition*. [S.l.], 2009. p. 248–255. Citado na página 30.

GAO, Y.; ZHOU, M.; LIU, D.; YAN, Z.; ZHANG, S.; METAXAS, D. N. *A Data-scalable Transformer for Medical Image Segmentation: Architecture, Model Efficiency, and Benchmark*. arXiv, 2022. Disponível em: <https://arxiv.org/abs/2203.00131>. Citado na página 15.

GOODFELLOW, I.; BENGIO, Y.; COURVILLE, A. *Deep Learning*. [S.l.: s.n.], 2016. Citado na página 18.

HASSANIEN, M. A.; SINGH, V. K.; PUIG, D.; ABDEL-NASSER, M. Predicting breast tumor malignancy using deep convnext radiomics and quality-based score pooling in ultrasound sequences. *Diagnostics*, v. 12, n. 5, 2022. ISSN 2075-4418. Disponível em: <https://www.mdpi.com/2075-4418/12/5/1053>. Citado na página 15.

KHALIL, H.; A.EL-HAG, N.; SEDIK, A.; EL-SHAFAI, W.; MOHAMED, A.; KHALAF, A. A. M.; BANBY, G. E.; EL-SAMIE, F. A.; EL-FISHAWY, A. Classification of diabetic retinopathy types based on convolution neural network (cnn). *Menoufia Journal of Electronic Engineering Research*, v. 28, p. 126–153, 12 2019. Citado na página 15.

LEMAîTRE, G.; NOGUEIRA, F.; ARIDAS, C. K. Imbalanced-learn: A python toolbox to tackle the curse of imbalanced datasets in machine learning. *Journal of Machine Learning Research*, v. 18, n. 17, p. 1–5, 2017. Disponível em: <http://jmlr.org/papers/v18/16-365.html>. Citado na página 32.

LI, J.; WANG, C.; HUANG, B.; ZHOU, Z. *ConvNeXt-backbone HoVerNet for nuclei segmentation and classification*. arXiv, 2022. Disponível em: <https://arxiv.org/abs/2202.13560>. Citado na página 16.

LI, L.; JAMIESON, K.; DESALVO, G.; ROSTAMIZADEH, A.; TALWALKAR, A. Hyperband: A novel bandit-based approach to hyperparameter optimization. *Journal of Machine Learning Research*, v. 18, n. 185, p. 1–52, 2018. Disponível em: <http://jmlr.org/papers/v18/16-558.html>. Citado na página 30.

LIU, Z.; LIN, Y.; CAO, Y.; HU, H.; WEI, Y.; ZHANG, Z.; LIN, S.; GUO, B. *Swin Transformer: Hierarchical Vision Transformer using Shifted Windows*. arXiv, 2021. Disponível em: <https://arxiv.org/abs/2103.14030>. Citado na página 22.

LIU, Z.; MAO, H.; WU, C.-Y.; FEICHTENHOFER, C.; DARRELL, T.; XIE, S. *A ConvNet for the 2020s*. arXiv, 2022. Disponível em: <https://arxiv.org/abs/2201.03545>. Citado 4 vezes nas páginas 14, 21, 23 e 30.

O'MALLEY, T.; BURSZTEIN, E.; LONG, J.; CHOLLET, F.; JIN, H.; INVERNIZZI, L. et al. *Keras Tuner*. 2019. <https://github.com/keras-team/keras-tuner>. Citado na página 30.

S.B C; STALMANS INGEBORGD; AHMED IQBAL IKE K.E; SNG, C. C. f. T. N. Y. F. D. Glaucoma screening: where are we and where do we need to go? *Current Opinion in Ophthalmology*, 2020. Citado na página 17.

TAN, M.; LE, Q. V. Efficientnet: Rethinking model scaling for convolutional neural networks. arXiv, 2019. Disponível em: <https://arxiv.org/abs/1905.11946>. Citado 4 vezes nas páginas 14, 19, 20 e 30.

TOUVRON, H.; CORD, M.; EL-NOUBY, A.; VERBEEK, J.; JéGOU, H. *Three things everyone should know about Vision Transformers*. arXiv, 2022. Disponível em: <https://arxiv.org/abs/2203.09795>. Citado na página 14.

UNIVERSITYIAI-PKU, I. of Artificial Intelligence at P. *ODIR-5k*. 2019. <https://odir2019.grand-challenge.org/dataset/>. Citado na página 26.

WANG, J.; YANG, L.; HUO, Z.; HE, W.; LUO, J. Multi-label classification of fundus images with efficientnet. *IEEE Access*, v. 8, p. 212499–212508, 01 2020. Citado 4 vezes nas páginas 15, 35, 36 e 37.

XIE, S.; GIRSHICK, R.; DOLLáR, P.; TU, Z.; HE, K. *Aggregated Residual Transformations for Deep Neural Networks*. arXiv, 2016. Disponível em: <https://arxiv.org/abs/1611.05431>. Citado na página 23.