NATANAEL LUCENA DE MEDEIROS

DETECÇÃO DA PSORÍASE UTILIZANDO VISÃO COMPUTACIONAL: UMA ABORDAGEM COMPARATIVA ENTRE CNNS E VISION TRANSFORMERS

Trabalho de Conclusão de Curso apresentado à banca avaliadora do Curso de Engenharia de Computação, da Escola Superior de Tecnologia, da Universidade do Estado do Amazonas, como pré-requisito para obtenção do título de Engenheiro de Computação.

Orientador(a): Prof. Dr. Fábio Santos da Silva

Manaus – Dezembro – 2024

Ficha Catalográfica

Ficha catalográfica elaborada automaticamente de acordo com os dados fornecidos pelo(a) autor(a).
Sistema Integrado de Bibliotecas da Universidade do Estado do Amazonas.

M488d Medeiros, Natanael Lucena de
Detecção da psoríase utilizando visão computacional : uma abordagem comparativa entre CNNs e Vision Transformers / Natanael Lucena de Medeiros . Manaus : [s.n], 2024.
89 f.: color.; 21,0 cm.

TCC - Graduação em Engenharia de Computação- Universidade do Estado do Amazonas, Manaus, 2024.
Inclui Bibliografia.
Orientador: Silva, Fábio Santos da.

1. Visão computacional. 2. Redes Neurais. 3. Aprendizado Profundo. 4. Transformers. 5. Dermatologia. I. Silva, Fábio Santos da (Orient.) II. Universidade do Estado do Amazonas. III. Título

CDU(1997)62:004

Universidade do Estado do Amazonas - UEA
Escola Superior de Tecnologia - EST
Curso de Engenharia da Computação

FOLHA DE APROVAÇÃO

Detecção da Psoríase Utilizando Visão Computacional: Uma Abordagem Comparativa entre CNNs e Vision Transformers.

Natanael Lucena de Medeiros

Trabalho de Conclusão de Curso defendido e aprovado pela banca avaliadora constituída pelos professores:

Fabio Santos da Silva
Presidente

Carlos Maurício Serodio Figueiredo
Avaliador(a)

Tiago Eugenio de Melo
Avaliador(a)

Manaus, 20 de Dezembro de 2024

Agradecimentos

A Deus, pela vida e por me guiar e sustentar ao longo de toda essa jornada; a minha mãe Dalzamir e meu pai Luciano, por todo o amor, amparo e cuidado durante toda a minha vida; à minha esposa Vitória por ser o meu incentivo; a meus colegas de graduação Gladson, Ademir e Ronald, pela amizade e apoio, especialmente nesse período de graduação; ao professor Ricardo Rios pelos ensinamentos dados nas salas de aula e no decorrer dos trabalhos desenvolvidos. A sua orientação, conhecimento e experiência foram de extrema importância para a construção desse trabalho e para a minha formação acadêmica; a esta universidade, seu corpo docente, direção e administração que propiciaram a realização do ensino superior; a todos que, de forma direta ou indireta, contribuíram para a concretização deste trabalho.

Resumo

Esse trabalho apresenta uma análise comparativa entre modelos de Rede Neurais Convolucionais (CNNs) e de *Vision Transformers* (ViT) no que diz respeito à tarefa de multiclassificação de imagens de pele com psoríase e similares, como, por exemplo, dermatite, línquen plano e pitiríase rosada. O objetivo do experimento é identificar o modelo mais indicado para a detecção de psoríase, como ferramenta de diagnóstico da doença. Na implementação, cada modelo é pré-treinado a partir do ImageNet e adaptado para o conjunto de imagens coletado. Durante a análise comparativa, observou-se que ambas as arquiteturas, CNN e ViT, apresentam ótimas métricas de predição nos experimentos, mas os ViTs se destacam por atingir um desempenho ligeiramente superior com modelos significativamente menores. Por fim, o DaViT-B (*Dual Attention Vision Transformer - Base*) se destacou entre todos, sendo, portanto, selecionado como o modelo mais indicado para a detecção da psoríase, apresentando um *f1-score* de 96,4%.

Palavras-Chave: Inteligência Artificial; Redes Neurais Convolucionais; *Transformers*; Visão Computacional; Psoríase.

Abstract

This work performs a comparative analysis between Convolutional Neural Networks (CNNs) and Vision Transformers (ViT) models on the task of multiclassifying skin images, detecting psoriasis among other images of skin diseases visually similar to it, such as dermatitis, lichen planus and pityriasis rosea. The aim of the experiment is to identify the most suitable model for detecting psoriasis, so that it can be used as a diagnostic tool for the disease. During implementation, each model is pre-trained from ImageNet and adapted to the collected image set. During the comparative analysis, it is observed that both CNN and ViT architectures present excellent prediction metrics in the experiment, but ViTs stand out by achieving slightly better performance with significantly smaller models. Finally, DaViT-B (Dual Attention Vision Transformer - Base) stood out from the rest and was therefore selected as the most suitable model for detecting psoriasis, with an f1-score of 96.4%.

Keywords: Artificial Intelligence; Convolutional Neural Networks; Transformers; Computer Vision; Psoriasis.

Sumário

Lista de Tabelas

Lista de Figuras

Lista de Abreviaturas e Siglas

Adam	*Adaptive Moment Estimation*
AI	*Artificial Intelligence*
ANN	*Artificial Neural Network*
CA	*Cross-Attention*
CNN	*Convolutional Neural Network*
DaViT	*Dual Attention Vision Transformer*
dHash	*Differencial Hash*
DNN	*Deep Neural Network*
ELU	*Exponential Linear Unit*
GELU	*Gaussian Error Linear Unit*
IA	Inteligência Artificial
ILSVRC	*ImageNet Large Scale Visual Recognition Challenge*
MaxViT	*Multi-Axis Vision Transformer*
ML	*Machine Learning*
MLP	*Multi-Layer Perceptron*
MSA	*Masked Self-Attention*
NLP	*Natural Language Processing*
PLN	Processamento de Linguagem Natural
ReLU	*Rectified Linear Unit*
RMSprop	*Root Mean Square Propagation*
SGD	*Stochastic Gradient Descent*

SVM *Support Vector Machine*

ViT *Vision Transformer*

XAI *Explainable AI*

Capítulo 1

Introdução

A psoríase é uma doença inflamatória da pele, recorrente e sem cura. Estima-se que 3% da população mundial, isto é, aproximadamente 125 milhões de pessoas, sofram com os sintomas da doença (DERMATOLOGIA, 2023). No Brasil, estimativas apontam que 90% dos entrevistados, de um total de 2.080 pessoas em 130 cidades, desconhecem a doença e apenas 6% reconhecem as lesões da psoríase (BRASIL, 2020).

Atualmente, o diagnóstico da psoríase é realizado por dermatologistas que se baseiam em achados clínicos e pistas visuais e hápticas (DASH et al., 2020). Durante a consulta, o profissional procura identificar a presença de sinais característicos da doença, examinando protuberâncias, erupções cutâneas, escamas e a descoloração da pele, e observando em que local elas aparecem na pele, diferenciando a psoríase de outras doenças com características similares (DAS, 2023). Biópsias podem ser necessárias para confirmar o diagnóstico (RODRIGUES; TEIXEIRA, 2009b). Esse procedimento é muito subjetivo, combinando observação e experiência prática do médico. Isso, comumente, impõe uma demora e maior complexidade para o diagnóstico conclusivo (JI YIYI WANG, 2022).

A aplicação de técnicas de visão computacional para a detecção e classificação de doenças de pele a partir de imagens apresenta-se como um método eficaz para a solução desses problemas. Essas técnicas oferecem a capacidade de explorar características da pele humana na imagem que não poderiam ser facilmente visíveis ao olho humano. Além disso, fornece uma solução objetiva e com potencial de automatização (GUIMARãES et al., 2018; MOURA ALESSANDRA

MARINS COELHO, 2021).

Uma das técnicas de visão computacional é a adoção de métodos de classificação baseados em Inteligência Artificial (IA), especificamente Aprendizado Profundo (*Deep Learning*), como suporte ao profissional médico. Tais métodos têm-se mostrado promissores no que diz respeito à eficácia e à eficiência do diagnóstico de doenças (OLESCKI, 2021; SILVA et al., 2022). Especificamente, os modelos das categorias *Vision Transformer* – ViT e Rede Neural Convolucional (*Convolutional Neural Network* – CNN) alcançaram um desempenho próximo ou até mesmo superior ao dos seres humanos no que se diz respeito à classificação, tratamento e análise de imagens (WEI et al., 2023; MOHAN et al., 2024; QUINTERO-ROJAS; GONZáLEZ, 2021; ROSLAN et al., 2020; WU et al., 2019; MADURANGA; NANDASENA, 2022).

Este trabalho apresenta um estudo comparativo de alguns modelos *Deep Learning* baseados na arquitetura de CNN e de outros baseados na arquitetura ViT. As comparações são realizadas utilizando uma tarefa de multi-classificação de imagens de pele humana com psoríase. Entre as imagens estão incluídas as de lesões de pele causadas por doenças visualmente similares às da psoríase. O intuito disso é o de promover a automatizaç ão de parte do processo de diagnóstico de psoríase, disponibilizando aos dermatologistas uma ferramenta de apoio para confirmação de hipóteses diagnósticas.

O trabalho de (MILANI et al., 2023) foi utilizado como base, sendo que o conjunto de dados com novas imagens foi obtido a partir de fontes atualizadas. Essa ampliação inclui imagens de doenças de pele similares à psoríase, como, por exemplo, dermatite, pitiríase rosada e líquen plano. As imagens duplicadas foram tratadas no conjunto.

Foram incluídos no experimento os modelos de CNN: EfficientNet-V2-L (TAN; LE, 2021) e ConvNeXt-L (LIU et al., 2022). Com isso, para o problema proposto nesse trabalho, compara-se o desempenho das CNNs selecionadas com o de modelos de ViT (*Transformers* especializados em tarefas de visão computacional): ViT-L/16 (ALEXEY, 2020), MaxViT-T (TU et al., 2022) e DaViT-B (DING et al., 2022). A seleção dos modelos foi realizada mediante pesquisa exploratória sobre a literatura científica recente e *benchmarks* atuais do ImageNet (DENG et al., 2009).

Este trabalho está estruturado em seis capítulos, incluindo a introdução. Na sequência da introdução, são apresentados o contexto e as justificativas para o estudo, os objetivos (geral e específicos), a metodologia aplicada, o cronograma de atividades e os materiais necessários. O Capítulo 2 reúne uma revisão da literatura pertinente ao contexto do trabalho. No Capítulo 3, é realizada uma análise dos trabalhos correlatos. O Capítulo 4 explora a proposta de métodos e implementação. O Capítulo 5 detalha os resultados obtidos e as discussões correspondentes. Finalmente, o Capítulo 6 apresenta as considerações finais do trabalho.

1.1 Justificativa

A psoríase foi identificada inicialmente na Grécia Antiga. Celsus (925 a.C. – 45 d.C.) fez a primeira descrição da doença. Em seguida, Hipócrates (946–375) descreveu as condições inflamatórias da pele que se assemelhavam à psoríase e as classificou denominando-as *lopoi* (descamar). Galeno (133 – 200 d.C.) foi o primeiro a usar o termo psoríase, do grego psora (prurido) e a identificá-la como uma condição de saúde de pele. Entretanto, pesquisadores e historiadores argumentam que, como Galeno descreveu uma erupção pruriginosa e escamosa das pálpebras e outras regiões, supõe-se que seu relato se referia ao eczema seborreico (ROMITI et al., 2009).

Apesar de ter sido identificada e tratada há vários milênios, a psoríase continua muito desconhecida. Estima-se que 3% da população mundial, ou seja, aproximadamente 240 milhões de pessoas, sofrem com os sintomas da psoríase (DERMATOLOGIA, 2023). No Brasil, estimativas apontam que 90% dos entrevistados, de um total de 2.080 pessoas em 130 cidades, desconhecem a doença e apenas 6% reconhecem as lesões da psoríase. Normalmente, as lesões causadas pela doença são confundidas com alergias, câncer de pele, hanseníase e micoses (BRASIL, 2020).

O equívoco na identificação da doença ocorre porque as lesões causadas pela psoríase possuem, considerando sua gravidade, tamanhos diferentes e formas arbitrárias (LIN et al., 2021). Os contornos das lesões são ambíguos, a diferença entre as lesões da psoríase e das outras doenças não é bem caracterizada e várias outras áreas da pele, acometidas pelas doenças, estão dentro do intervalo de observação da psoríase (MOON et al., 2021). Essa semelhança com

as demais doenças de pele dificulta a identificação do tipo de anormalidade (ROSLAN et al., 2020).

Os acometidos de psoríase apresentam outras enfermidades, como, por exemplo, doenças cardiovasculares, doença inflamatória intestinal, obesidade, ansiedade e depressão, para citar algumas delas (DERMATOLOGIA, 2016). A detecção precoce, portanto, é imprescindível para promover o bem-estar e uma melhor qualidade de vida dos pacientes com psoríase (QUINTERO-ROJAS; GONZáLEZ, 2021).

O diagnóstico da psoríase é realizado atualmente por dermatologistas que se baseiam em achados clínicos, pistas visuais e hápticas (DASH et al., 2020). Biópsias podem ser realizadas para confirmar o diagnóstico (RODRIGUES; TEIXEIRA, 2009b). São verificadas na consulta a presença ou ausência de sinais, morfologia, distribuição, coloração, descamação e disposição das lesões. O procedimento adotado atualmente para a identificação da doença é a combinação de observação e experiência médica. Isso, como pode ser deduzido, é consideravelmente subjetivo e torna o diagnóstico conclusivo demorado e complexo (JI YIYI WANG, 2022). Uma abordagem para tornar o método mais eficiente e eficaz é a utilização de métodos de classificação baseados em Inteligência Artificial (IA).

1.2 Objetivos

Este trabalho tem por objetivo geral comparar o desempenho de algumas CNNs (EfficientNet-V2-L (TAN; LE, 2021), ConvNeXt-L (LIU et al., 2022) e Inception-v3 (SZEGEDY et al., 2016b)) e ViTs (ViT-L/16 (ALEXEY, 2020), MaxViT-T (TU et al., 2022) e DaViT-B (DING et al., 2022)) na tarefa de multi-classificação de imagens de lesões cutâneas causadas pela psoríase. Em particular, busca-se selecionar o modelo mais promissor para atuar como ferramenta de diagnóstico dermatológico.

1.2.1 Objetivos Específicos

Para alcançar o objetivo geral, os seguintes objetivos específicos são necessários:

- Realizar um levantamento bibliográfico sobre a psoríase e sobre as arquiteturas CNN e ViT aplicadas à tarefas de classificação de imagens;
- Coletar e tratar uma base de dados de imagens de pele com psoríase, outras doenças com características visualmente similares às da psoríase, e de pele saudável;
- Comparar os desempenhos das arquiteturas CNNs e ViTs com base em métricas de avaliação previamente definidas.
- Avaliar o desempenho individual dos modelos selecionados, identificando o modelo mais adequado à tarefa proposta.

1.3 Metodologia

Esse trabalho possui caráter aplicado e busca realizar uma investigação exploratória com base em materiais bibliográficos e experimentais. A coleta de dados foi conduzida por meio de observação direta e análise de registros documentais. Os dados obtidos foram examinados utilizando uma abordagem quantitativa.

A etapa inicial desse trabalho envolveu a busca por bases de dados contendo imagens de peles, tanto com lesões de psoríase e doenças similares quanto sem. Após essa coleta, os dados foram analisados com o objetivo de identificar métodos de pré-processamento que pudessem otimizar o treinamento da rede neural.

Após a etapa inicial, foi conduzida uma análise das arquiteturas de CNNs e ViTs mais aplicadas na classificação de imagens, com foco em identificar modelos que apresentassem bom desempenho e estivessem alinhados aos objetivos do projeto. Para tanto, foi realizada uma revisão aprofundada da literatura e pesquisas em *benchmarks* populares.

Na etapa seguinte, as arquiteturas selecionadas foram implementadas, utilizando uma base de imagens previamente dividida em conjuntos de treinamento, validação e teste. Os algoritmos foram treinados e validados, com a aplicação de estratégias para otimizar o processo de aprendizado. O conjunto de validação serviu para avaliar o desempenho dos modelos e inter-

pretar seus resultados. Assim, foi possível identificar aspectos a serem aprimorados e ajustar os parâmetros da rede neural, permitindo o retreinamento dos modelos, caso necessário.

Foi indispensável definir métricas de avaliação para interpretar os resultados obtidos a partir do conjunto de teste e realizar comparações entre os modelos. Por fim, foi realizada a redação do trabalho monográfico.

1.4 Materiais

- Hardware: CPU Intel(R) Core(TM) i7-8700, 2 GPU NVIDIA GeForce GTX 1080 TI, RAM 56 GB;
- Base de dados obtida a partir de plataformas dermatológicas certificadas na Web;
- Linguagem Python 3.12.3 (FOUNDATION, 2024): A linguagem Python foi escolhida pois é amplamente utilizada em trabalhos da área de Inteligência Artificial, apresentando como uma de suas principais vantagens o amplo suporte às bibliotecas desse escopo. Por isso, a mesma será utilizada nesse trabalho para o desenvolvimento de cada *script* que gerou o modelo classificador;
- Biblioteca Pytorch 2.3.0 (PASZKE et al., 2019): A biblioteca Pytorch é amplamente utilizada em pesquisas da área de Inteligência Artificial, pois oferece uma infraestrutura flexível e eficiente. É poderosa para, por exemplo, a criação e treinamento de redes neurais profundas. A biblioteca, portanto, foi utilizada como base para o pré-processamento das imagens, para o ajuste fino, treinamento, e validação de cada modelo;
- Outras Bibliotecas do Python: Scikit-learn (PEDREGOSA et al., 2011), Pandas (MCKINNEY, 2010), Numpy (HARRIS et al., 2020), Matplotlib (HUNTER, 2007), TorchVision (MAINTAINERS; CONTRIBUTORS, 2016), Timm (WIGHTMAN, 2019), Seaborn (WASKOM, 2021).

Capítulo 2

Fundamentação Teórica

Nesse capítulo são apresentados os principais conceitos que permeiam esse trabalho. As seções abordam a doença de pele psoríase, e os conceitos de Inteligência Artificial, Aprendizado de Máquina, Aprendizado Supervisionado, Aprendizado Profundo, Redes Neurais Artificiais, Redes Neurais Convolucionais, *Vision Transformers* e, por fim, Métricas de Avaliação.

2.1 Psoríase

A psoríase é uma doença cutânea, desfigurante, inflamatória e crônica, e de escala global. Dados indicam que a prevalência da psoríase varia de acordo com a localização geográfica (Figura 2.1), sendo a maior prevalência da doença na Dinamarca (1,91%) e a menor em Taiwan (0,06%) (ATLAS, 2024). Outro dado global dessa mesma fonte é o de que 81% dos países do mundo carecem de informações sobre a epidemiologia da psoríase.

A causa da psoríase ainda é desconhecida, embora fatores ambientais e genéticos se apresentem como indicativos importantes. Apesar disso, há evidências de que existe um componente genético no progresso da doença; muitos fatores ambientais têm sido associados à psoríase e estiveram envolvidos na indução do progresso da doença, resultando em um quadro de piora dos pacientes. Esses fatores incluem: trauma físico (EYRE; KRUEGER, 1982); infecções (LAZAR; ROENIGK, 1988); estresse (SEVILLE, 1977); alguns medicamentos, como, por exemplo, betabloqueadores, lítio, antimaláricos e esteroides sistêmicos (BASAVARAJ et al., 2010); hipo-

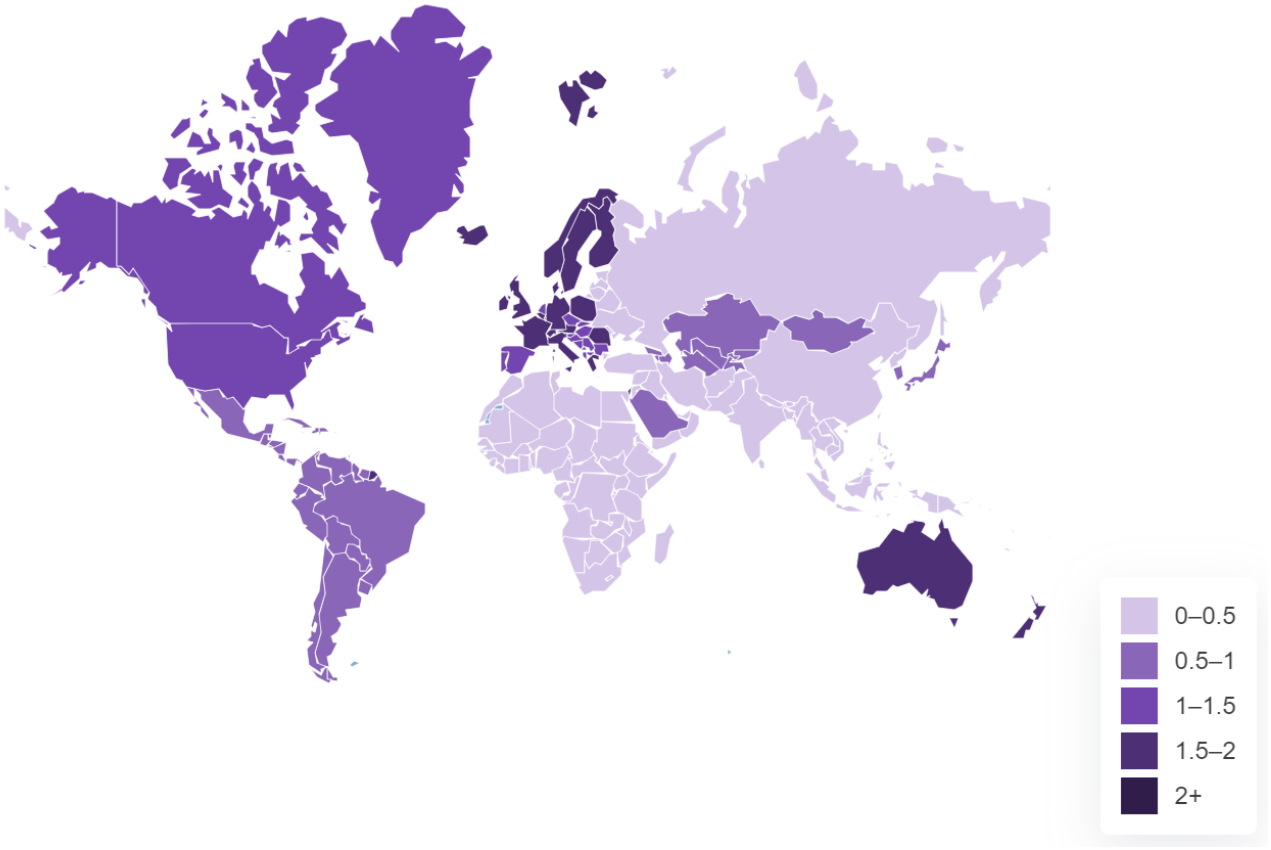

Figura 2.1: Percentual de prevalência da psoríase em escala global (ATLAS, 2024).

calcemia (STEWART JULIA BATTAGLINI-SABETTA, 1984); consumo de álcool, tabagismo (JANKOVI et al., 2009); e clima (BALATO et al., 2013).

Conforme mencionado, existem muitas evidências de que a psoríase tem um fator genético relevante, uma vez que observou-se que a doença tendia a ocorrer entre membros da mesma família. Dados robustos apoiando uma base genética para a psoríase vêm de estudos que examinam a concordância para a doença em gêmeos (BRANDRUP et al., 1982).

O diagnóstico da psoríase é principalmente clínico (erupções cutâneas, alterações ungueais e implicações articulares). Existem diferentes tipos clínicos de psoríase, sendo o mais comum a psoríase em placas (Figura 2.2), que afeta a maioria dos pacientes (PAPP et al., 2011).

Embora a psoríase congênita seja muito rara, sua primeira manifestação pode ocorrer em qualquer idade, porém raramente se manifesta antes dos 10 anos de idade. A maioria das formas de psoríase está presente antes dos 30 anos de idade (HENSELER, 1985). Cronicidade, inflamação e hiperproliferação são as características principais da psoríase na infância (DHAR

Figura 2.2: Descamação da pele do antebraço, causada pela psoríase em placas (DERMATOLOGY, 2024).

et al., 2011).

A psoríase é uma doença ainda sem cura. Há períodos que apresenta remissões e outros exacerbações. Cerca de 10% dos casos da doença evoluem para artrite e estimativas apontam que homens e mulheres com psoríase grave morrem 3,5 e 4,4 anos antes, respectivamente, em comparação com homens e mulheres sem a doença (GELFAND et al., 2007). Estudos populacionais apontam que pacientes com psoríase possuem maior risco de adquirir outras comorbidades médicas, como, por exemplo, doenças cardiovasculares, doenças pulmonares crônicas, diabetes mellitus, doença renal, doença de Crohn, penfigoide bolhoso, vitiligo e problemas articulares (YEUNG et al., 2013).

2.1.1 Doenças Visualmente Similares à Psoríase

A psoríase pertence ao grupo de doenças chamado *Doenças Descamativas*, que também inclui distúrbios como dermatites, parapsoríase, pitiríase rosada, pitiríase rubra pilar, líquen plano e líquen escleroso. As protuberâncias, erupções cutâneas, escamas e descoloração da pele dessas doenças apresentam características semelhantes às da psoríase (DAS, 2023). Para esse trabalho, foram coletadas imagens de pele com lesões causadas pelas seguintes doenças similares à psoríase: pitiríase rosada; dermatites; e líquen plano.

Dermatite é um termo abrangente que descreve inflamações da pele, como dermatite atópica e dermatite de contato. Caracteriza-se por erupções cutâneas, coceira e, em casos graves, descamação e fissuras. Assim como a psoríase, as dermatites apresentam inflamação como base fisiopatológica comum (CHOVATIYA; SILVERBERG, 2019). No entanto, diferentemente da psoríase, estudos indicam que a desregulação imunológica é um ponto de interseção quanto à causa da doença (SILVERBERG, 2017).

Líquen plano é uma doença inflamatória crônica de origem autoimune que afeta pele e mucosas e apresenta lesões arroxeadas e pruriginosas. Sua semelhança com a psoríase, além do aspecto visual das lesões, reside no envolvimento do sistema imune adaptativo, com proliferação de linfócitos T. (SOUZA et al., 2016).

Pitiríase rosada é uma doença eruptiva cutânea autolimitada, benigna, e de provável origem viral. Apesar de não ser crônica como a psoríase, também apresenta placas semelhantes (SILVA et al.,). Pesquisas enfatizam a necessidade da dermatoscopia para diferenciar condições semelhantes, incluindo a psoríase (RODRIGUES; TEIXEIRA, 2009a).

A semelhança visual dessas doenças com a psoríase fomenta o desafio de oferecer um diagnóstico precoce mais objetivo e preciso da doença. A Figura 2.3 exemplifica essa semelhança apresentando amostras das lesões de pele causadas por cada uma dessas doenças.

Apresentando-se como possível solução a esse impasse, ferramentas baseadas em Visão Computacional podem ter o potencial de mitigar os erros clínicos, pois se baseiam na análise humana subjetiva sobre as características visuais das lesões na pele. Nesse sentido, métodos de Inteligência Artificial propostos para a resolução de problemas de Visão Computacional tornam-se soluções interessantes, modernas e com potencial de eficácia. Nas seções 2.2 a 2.9 serão abordados os conceitos envolvendo a área de Inteligência Artificial e suas ramificações relevantes para este trabalho.

2.2 Inteligência Artificial

Existem diversas definições para Inteligência Artificial (IA) (*Artificial Intelligence*, em inglês – AI), como por exemplo: sistemas que pensam como humanos, que agem como humanos,

(a) Psoríase (b) Dermatite

(c) Líquen plano (d) Pitiríase rosada

Figura 2.3: Aspecto visual da psoríase (a) em comparação com doenças similares (b), (c) e (d) (DERMNETNZ, 2024).

que raciocinam de maneira lógica, ou que agem racionalmente (RUSSELL; NORVIG, 2009). Essa abrangência ocorre porque os conceitos de inteligência e comportamento inteligente são dinâmicos e adaptáveis (COZMAN; PLONSKI; NERI, 2021). De forma geral, a IA pode ser entendida como uma área da Ciência da Computação que se dedica ao desenvolvimento de métodos e ferramentas computacionais que imitam o comportamento inteligente (HAMMOND, 2015). Assim, atividades tradicionalmente realizadas por humanos podem ser executadas por máquinas com um nível de autonomia satisfatório.

Entre suas principais aplicações, a IA tem ganhado espaço em áreas como reconhecimento de linguagem natural (BRAGA et al., 2021) e reconhecimento de imagens (LI, 2022).

2.3 Aprendizado de Máquina

O Aprendizado de Máquina (*Machine Learning*, em inglês – ML) é uma área da IA que permite que computadores executem tarefas específicas sem depender de uma programação detalhada para cada ação (HURWITZ; KIRSCH, 2018).

Um exemplo clássico de aplicação de ML são os sistemas de recomendação. Esses sistemas utilizam algoritmos que combinam informações para analisar dados sobre o comportamento e preferências dos usuários, sugerindo itens personalizados para cada perfil (BURKE, 2002). São amplamente empregados em plataformas de *streaming* de vídeo e música, comércio eletrônico, redes sociais, entre outras.

O ML baseia-se em diversos algoritmos que, ao aprenderem iterativamente a partir de dados, conseguem gerar e refinar resultados. Quanto mais dados são usados no treinamento desses algoritmos, maior a chance de recomendações precisas, pois o volume de dados contribui para análises mais detalhadas (KOWALCZYK; SZLÁVIK; SCHUT, 2011). No entanto, o aumento do volume de dados, por si só, não garante bons resultados. É essencial que a modelagem e o tratamento dos dados, assim como a escolha dos métodos preditivos, sejam apropriados para o contexto do problema (ZHOU; KHEMMARAT; GAO, 2010).

Os algoritmos de ML podem ser divididos em: aprendizado supervisionado (Seção 2.4, não supervisionado, semi-supervisionado e por reforço. Nessa monografia, será dada maior atenção ao aprendizado supervisionado.

2.4 Aprendizado Supervisionado

No aprendizado supervisionado os dados utilizados como entrada possuem rótulos, o que significa que a saída correspondente a cada dado também é conhecida. Esse tipo de aprendizado visa identificar padrões nas classificações dos dados, permitindo que uma função mapeie a entrada para a saída correta em processos futuros (BRINK; RICHARDS; FETHEROLF, 2016). Um exemplo prático desse método é o filtro de *spam*: e-mails são rotulados pelo usuário como "*spam*"ou "não *spam*", e o sistema aprende a classificar novos e-mails com base nesses exemplos

(GÉRON, 2019).

Os modelos supervisionados são amplamente aplicados em problemas diversos, como detecção de fraudes, reconhecimento de fala e análise de risco (HURWITZ; KIRSCH, 2018). Esse tipo de aprendizado busca comparar entradas e saídas dos dados históricos de treinamento para descobrir padrões e associações. Para isso, uma variedade de algoritmos e técnicas é utilizada, incluindo redes neurais artificiais, classificadores *Naive Bayes*, máquinas de vetores de suporte (*Support Vector Machine* – SVM) e florestas aleatórias (GÉRON, 2019). O método de redes neurais artificiais será explorado na Seção 2.5, pois é especialmente relevante para o desenvolvimento deste trabalho.

2.5 Redes Neurais Artificiais

O cérebro humano é composto por cerca de dez bilhões de células conhecidas como neurônios (ABRAHAM, 2005). Cada neurônio é constituído por três partes principais: o soma (ou corpo celular); um axônio; e múltiplos dendritos. O soma contém as organelas típicas de células encontradas no corpo humano e desempenha funções vitais para a sobrevivência celular. O axônio, por sua vez, é uma fibra longa que se estende a partir do soma, sendo responsável por transmitir sinais de um neurônio para outro, atuando como a saída neural. Os dendritos são fibras ramificadas que recebem sinais de outros neurônios e os conduzem ao soma, desempenhando o papel de entradas neurais.

A comunicação entre os neurônios ocorre por meio de sinapses, um processo em que os impulsos elétricos são transmitidos do axônio de um neurônio para os dendritos de outro (GURNEY, 1997). Essa troca de informações por meio de sinapses, organizada em uma vasta rede neural, é o mecanismo fundamental pelo qual o cérebro humano realiza suas funções (ABRAHAM, 2005). A Figura 2.4 ilustra os principais componentes presentes na estrutura de um neurônio biológico.

As Redes Neurais Artificiais (*Artificial Neural Networks*, em inglês – ANN) são modelos computacionais utilizados em aprendizado supervisionado que se inspiram no funcionamento das redes neurais biológicas. Essas redes são projetadas para adquirir conhecimento por meio

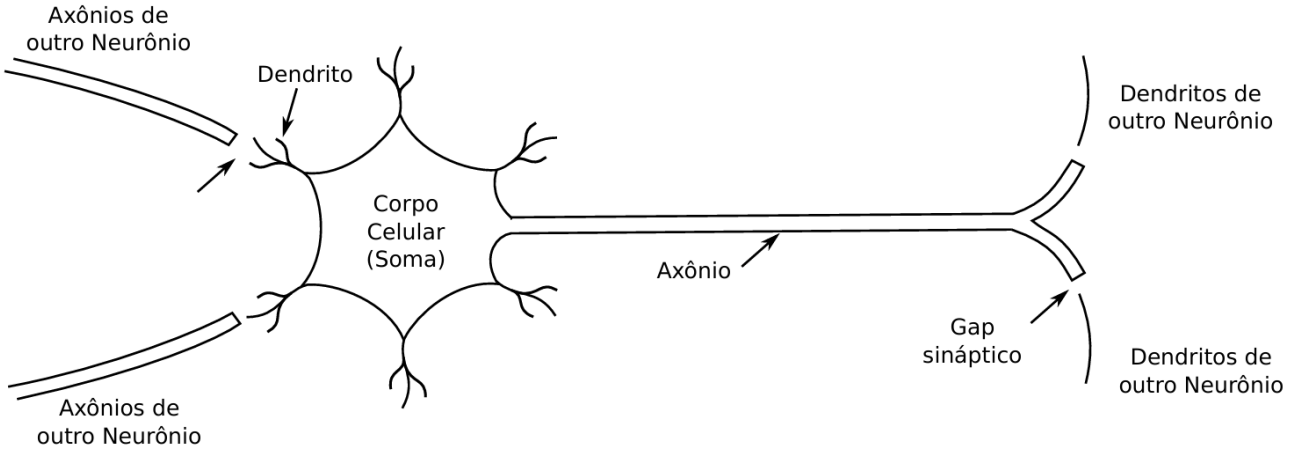

Figura 2.4: Estrutura básica do neurônio biológico (SILVA, 2014).

da experiência. A Figura 2.5 apresenta uma estrutura básica de uma ANN, conhecida como neurônio artificial.

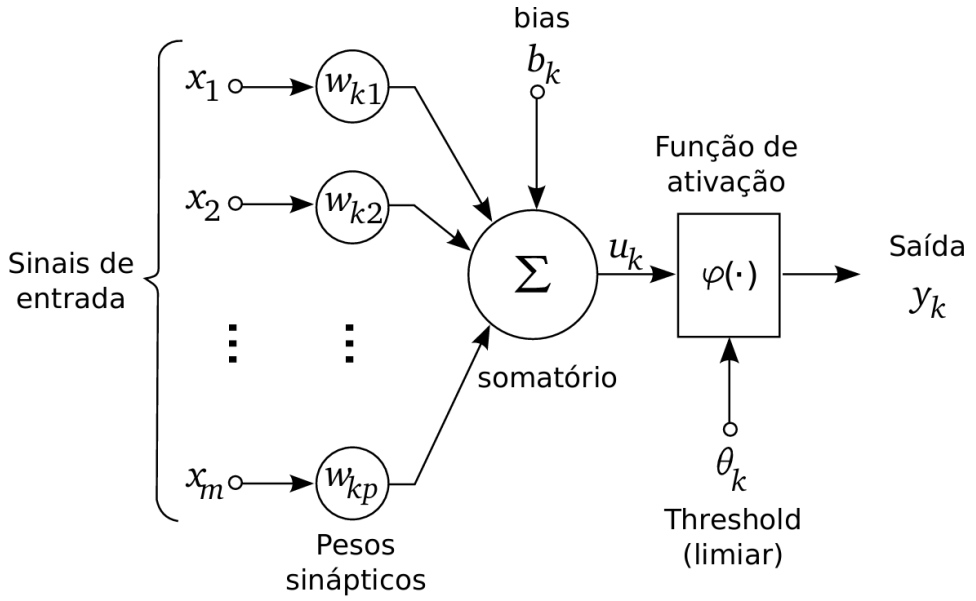

Figura 2.5: Modelo de Neurônio Artificial (SILVA, 2014).

O neurônio artificial é representado matematicamente com base nos princípios de um neurônio biológico. Segundo (HAYKIN, 1998), ele pode ser descrito a partir de três componentes principais:

- Conjunto de sinapses (sinais de entrada): Cada sinapse é associada a um peso que influencia a entrada correspondente. Para uma sinapse j, o valor de entrada X_j conectado ao neurônio k é multiplicado pelo peso sináptico, gerando um valor W_{kj}.

- Somador: Responsável por calcular a soma dos sinais de entrada, já ajustados pelos respectivos pesos sinápticos do neurônio.

- Função de ativação: Controla se o neurônio será ativado ou não. O resultado dessa função determina a saída final do neurônio.

No modelo ilustrado na Figura 2.5, o somador também recebe um viés (*bias*, em inglês), representado por b_k, como um sinal adicional. Esse viés, ajustável durante o processo de aprendizado, desempenha um papel importante ao aumentar a flexibilidade da rede, reduzindo a dependência do neurônio em relação às entradas (ROSENBLATT, 1958). Assim, a representação matemática de um neurônio artificial pode ser expressa pela Equação 2.1, onde y indica a saída do neurônio, x é o vetor de entrada, w é o vetor de pesos sinápticos, e b é o valor do viés.

$$y_k = \varphi\left(\sum_{i=1}^{m} x_i w_{ki} + b_k\right) \tag{2.1}$$

Os pesos e vieses compõem os parâmetros de uma ANN. O custo computacional de uma ANN é proporcional ao número de parâmetros e à quantidade de operações necessárias. Esse custo é discutido no artigo (LECUN; DENKER; SOLLA, 1989), que introduziu o conceito de poda de redes neurais para reduzir a complexidade sem comprometer o desempenho. Essa técnica é um exemplo de como o número de parâmetros impacta no custo computacional e na eficiência de predição.

2.5.1 Funções de Ativação

Diferentes tipos de funções de ativação podem atuar no neurônio artificial, desempenhando funções específicas na rede neural. Apresentam-se como as principais funções:

- Função Limiar: é uma das funções de ativação mais básicas e foi utilizada inicialmente nos perceptrons (ROSENBLATT, 1958), os primeiros modelos de redes neurais artificiais. Ela opera como uma função degrau, onde a saída é ativada com 1 se o valor de entrada exceder um determinado limiar e 0 caso contrário. Matematicamente, pode ser representada na Equação 2.2, onde x é a entrada e μ o limiar:

$$f(x) = \begin{cases} 1 & \text{se } x \geq \mu \\ 0 & \text{caso contrário} \end{cases} \tag{2.2}$$

- Função Linear: é simplesmente definida pela Equação 2.3, onde a e b são constantes. Ela é amplamente utilizada em camadas de saída de redes neurais quando o objetivo é prever valores contínuos, como em problemas de regressão. A principal característica da função linear é que ela mantém a proporcionalidade entre entrada e saída, o que é útil em modelos que não exigem não-linearidade (GOODFELLOW; BENGIO; COURVILLE, 2016).

$$f(x) = ax + b \tag{2.3}$$

- Sigmoide: A função sigmoide transforma os valores de entrada em um intervalo entre 0 e 1, e é representada pela equação 2.4. Sua principal aplicação está em problemas de classificação binária, onde as saídas podem ser interpretadas como probabilidades. Apesar disso, a sigmoide sofre de limitações, como, por exemplo, o problema do gradiente desvanecido, que dificulta o treinamento em redes profundas, e a ausência de centralidade em zero, o que pode desacelerar o aprendizado (GOODFELLOW; BENGIO; COURVILLE, 2016).

$$f(x) = \frac{1}{1 + e^{-x}} \quad 0 \leq x \leq 1 \tag{2.4}$$

- Tangente hiperbólica (tanh): é semelhante à sigmoide, mas mapeia os valores de entrada para o intervalo $[-1, 1]$. Isso a torna centrada em zero, o que promove um aprendizado mais equilibrado entre ativações positivas e negativas. Ainda que tenha vantagens sobre a sigmoide, a tangente hiperbólica também sofre do problema de gradiente desvanecido em valores extremos, o que pode limitar sua aplicação em redes muito profundas (ORR; MÜLLER, 1998). A função tangente hiperbólica é definida pela Equação 2.5.

$$f(x) = \frac{\sinh x}{\cosh x} = \frac{e^x - e^{-x}}{e^x + e^{-x}} \tag{2.5}$$

- ReLU: Unidade Linear Retificada (do inglês *Rectified Linear Unit*), a ReLU, definida pela equação 2.6, ativa apenas valores de entrada positivos, enquanto mapeia valores negativos para zero. Essa característica reduz a complexidade do cálculo e mitiga o problema do gradiente desvanecido em grande parte do domínio. Contudo, a função pode levar à chamada *morte de neurônios*, um problema em que as unidades param de responder permanentemente, resultando em gradientes zero e deixando de contribuir para o aprendizado (NAIR; HINTON, 2010).

$$f(x) = \max\{0, x\} \tag{2.6}$$

- Leaky ReLU: foi desenvolvida como uma solução ao problema da morte de neurônios na ReLU. A função assemelha-se com a da ReLU, com a diferença de que a Leaky ReLU permite que valores negativos tenham uma inclinação pequena, controlada por um hiperparâmetro α (vide Equação 2.7). Essa modificação reduz o risco de morte de neurônios e melhora a robustez do modelo (MAAS et al., 2013).

$$f(x) = \begin{cases} x & \text{se } x > 0 \\ \alpha x & \text{caso contrário} \end{cases} \tag{2.7}$$

- ELU: Unidade Linear Exponencial (do inglês *Exponential Linear Unit*) é uma variante da ReLU projetada para suavizar as ativações negativas com uma curva exponencial. Essa suavidade permite que a função corrija o viés de ativação média em valores não negativos, facilitando o aprendizado de pesos negativos e melhorando a robustez em tarefas complexas (CLEVERT; UNTERTHINER; HOCHREITER, 2020). No entanto, seu custo computacional é maior devido ao cálculo exponencial (vide Equação 2.8).

$$f(x) = \begin{cases} x & \text{se } x > 0 \\ \alpha\,(e^x - 1) & \text{caso contrário} \end{cases} \tag{2.8}$$

- Softmax: definida pela Equação 2.9 (j denota o número de classes), é amplamente utilizada em tarefas de multi-classificação. Ela converte as saídas da rede em probabilidades normalizadas, assegurando que a soma de todas as saídas seja igual a 1. Essa característica é crucial para interpretar as probabilidades associadas a cada classe, especialmente em tarefas como reconhecimento de imagens e processamento de linguagem natural.

$$f(z_i) = \frac{e^{z_i - \max(z)}}{\sum_{j=1} e^{z_j - \max(z)}} \tag{2.9}$$

- GELU: Unidade Linear de Erro Gaussiano (do inglês *Gaussian Error Linear Unit*), é uma função de ativação moderna que suaviza a ReLU utilizando uma função probabilística baseada na distribuição gaussiana. Devido o seu desempenho superior em tarefas de visão computacional e processamento de linguagem natural, a GELU é amplamente adotada em arquiteturas avançadas, como os modelos *Transformer* (HENDRYCKS; GIMPEL, 2016). A função é definida pela Equação 2.10, onde P é a função cumulativa da distribuição normal.

$$f(x) = x \times P(X \leq x) \tag{2.10}$$

2.5.2 *Multi-layer Perceptron*

Um único neurônio não é capaz de resolver problemas complexos, mesmo quando são usadas funções de ativação. Como ilustrado na Figura 2.6, é possível criar um hiperplano, permitindo a separação dos dados em categorias linearmente separáveis, como ocorre com operações booleanas como *AND*, *OR* e *NOT*. No entanto, surgem dificuldades ao lidar com problemas que não podem ser separados linearmente, como o caso do *XOR* (GRUS, 2016).

No mundo real, a grande maioria dos problemas não é linearmente separável e, para esses

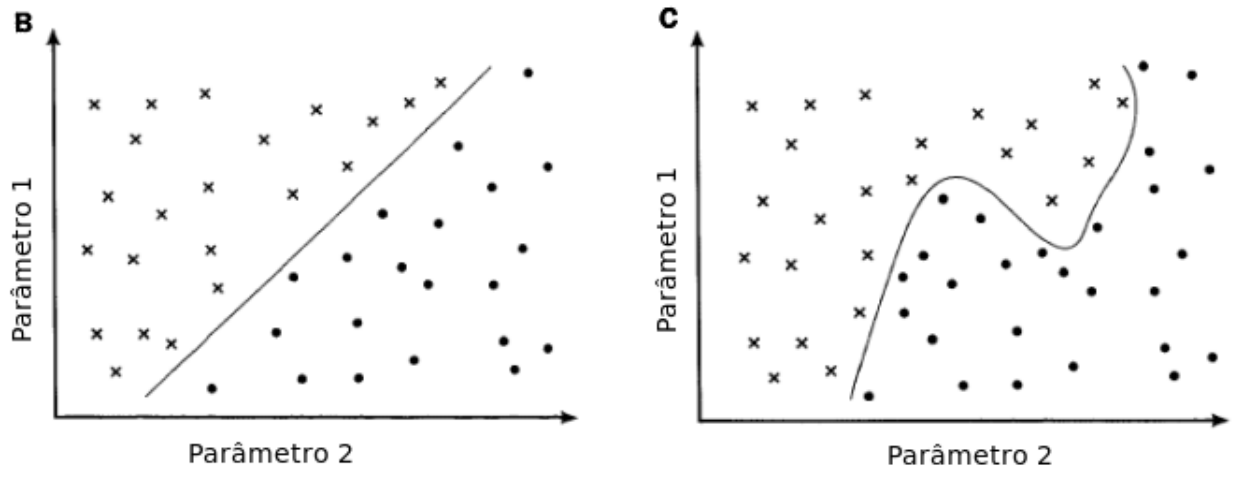

Figura 2.6: Problemas linearmente separáveis (B) e não linearmente separáveis (C) (GONÇALVES, 2015).

problemas, faz-se necessária a ligação de dezenas de neurônios sobre várias camadas de uma rede. Os modelos que possuem esse tipo de estrutura são definidos como **Perceptron Multicamadas** (*Multi-layer Perceptron*, em inglês – MLP).

Um MLP é um tipo de ANN organizado em três camadas principais: entrada, oculta e saída. A camada de entrada, composta por unidades, é responsável por converter os padrões apresentados à rede em valores numéricos. Na camada oculta, que pode conter uma ou mais subcamadas, ocorre o processamento mais intenso, onde são extraídas as características relevantes dos dados. Por fim, a camada de saída, também formada por unidades, recebe os resultados da camada oculta e gera um valor final que representa o dado processado (SOUZE, 2018). A Figura 2.7 ilustra a organização dessas camadas na estrutura desse tipo de rede.

Dessa forma, o MLP se destaca pela capacidade de resolver problemas que não podem ser separados linearmente. Por isso, suas aplicações mais comuns incluem tarefas como reconhecimento de fala, identificação de imagens e tradução automática entre idiomas (WASSERMAN; SCHWARTZ, 1988).

O aprendizado em uma MLP ocorre por meio de um algoritmo conhecido como *Backpropagation*, que é dividido em duas fases: *Forward Propagation* e *Backward Propagation* (RUMELHART; HINTON; WILLIAMS, 1986).

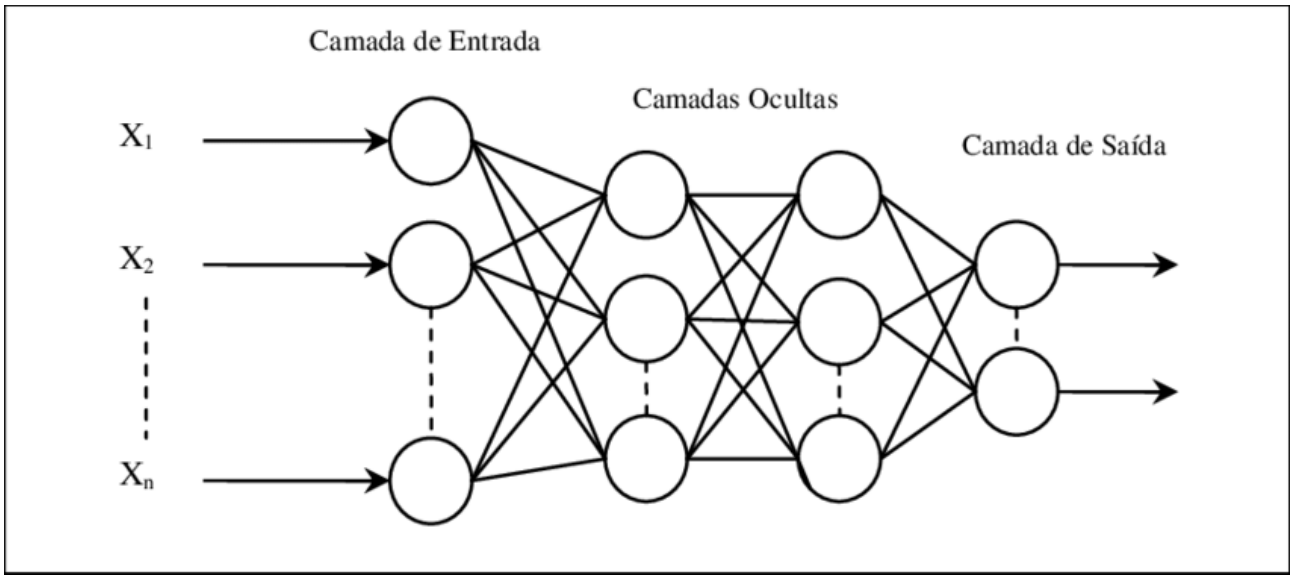

Figura 2.7: Perceptron Multicamada (SOBREIRO et al., 2008).

Na etapa de *Forward Propagation*, os dados são enviados para a camada de entrada, que os repassa progressivamente para as camadas seguintes, até chegar à saída final da rede. Durante esse percurso, os neurônios processam as informações, geralmente de maneira não linear, e as transmitem para a próxima camada. Esse processo se repete em todas as camadas da rede até que os resultados alcancem a camada de saída.

Inicialmente, a diferença entre a saída gerada pela rede e a saída esperada, chamada de erro de saída, tende a ser elevada, pois os pesos sinápticos ainda não estão ajustados. Para corrigir isso, inicia-se a fase de *Backward Propagation*, onde ocorre o ajuste efetivo dos pesos.

Nessa segunda etapa, o erro é propagado de volta através da rede, utilizando o método de descida do gradiente para encontrar os pesos adequados. Dessa forma, o ajuste é realizado de forma iterativa para cada peso em cada conexão sináptica (RUMELHART; HINTON; WILLIAMS, 1986):

1. Calcula-se o erro por meio da diferença entre o resultado obtido na saída da rede e o resultado esperado;

2. O erro é então multiplicado pela derivada da função de ativação, determinando o valor de variação;

3. Essa variação é propagada para trás na rede, sendo multiplicada pela derivada da função

de ativação de cada camada anterior, até alcançar a camada de entrada;

4. Após determinar todas as variações, os pesos são atualizados multiplicando a variação pela taxa de aprendizado.

O processo descrito é repetido até que seja alcançado o número de iterações estabelecido, ou até que o erro atinja o valor desejado. Usualmente, utiliza-se a métrica de épocas para estabelecer esse limite. Nesse contexto, o termo época refere-se ao momento em que todo o conjunto de dados passa pela rede ao menos uma vez. Além disso, o número de épocas necessário para treinar o modelo pode ser ajustado automaticamente.

Outro importante parâmetro de treinamento é o tamanho de lote (*batch size*), que define a quantidade de amostras processadas pela rede em cada iteração. Por exemplo, considerando um conjunto de dados com 1.000 amostras e um *batch size* de 100, o algoritmo processará as 100 primeiras amostras no treinamento inicial, depois as próximas 100, e assim por diante, até que todas as amostras tenham sido utilizadas.

Com exceção da camada de saída do MLP, todas as outras camadas possuem um neurônio de viés e estão completamente conectadas à camada subsequente. Isso significa que cada neurônio de uma camada está ligado a todos os neurônios da próxima. Essencialmente, Redes Neurais Profundas (*Deep Neural Networks*, em inglês – DNN) são um tipo de ANN caracterizado por conter duas ou mais camadas ocultas (GÉRON, 2019).

2.5.3 Otimizadores

Os otimizadores são fundamentais para o treinamento de redes neurais, ajustando iterativamente os pesos do modelo para minimizar a função de perda e melhorar a precisão nas previsões (DESAI, 2020). O papel do otimizador impacta diretamente a taxa de convergência, a estabilidade do treinamento e a capacidade do modelo de alcançar um ótimo global. Além disso, a escolha de um otimizador apropriado deve considerar o problema específico e a estrutura do modelo para garantir um desempenho eficiente e confiável (DESAI, 2020).

A seguir, serão brevemente apresentados alguns otimizadores populares, amplamente utilizados na literatura:

- SGD (ROBBINS; MONRO, 1951): abreviação de *Stochastic Gradient Descent*, é um método de otimização básico conhecido por sua simplicidade e eficiência. Ele atualiza iterativamente os parâmetros do modelo usando gradientes calculados em subconjuntos aleatórios do conjunto de dados, o que o torna computacionalmente barato e escalável para grandes volumes de dados (SHANTHI; MANIMEKALAI, 2024). Apesar de ser simples e eficiente, o SGD apresenta desvantagens significativas, como sua suscetibilidade a oscilações durante a convergência, especialmente em regiões com gradientes ruidosos, e a dificuldade de ajustar hiperparâmetros como a taxa de aprendizado (ASKARIZADEH et al., 2024).

- RMSprop (RUDER, 2016): abreviação de *Root Mean Square Propagation*, é um otimizador adaptativo projetado para lidar com taxas de aprendizado variáveis e estabilizar o treinamento de modelos de aprendizado profundo. Ele ajusta dinamicamente a taxa de aprendizado com base na média ponderada dos quadrados dos gradientes, permitindo que o treinamento seja mais robusto em superfícies de perda ruidosas (ELSHAMY et al., 2023). Apesar de ser capaz de convergir rapidamente e lidar bem com problemas não estacionários, o RMSprop depende do ajuste fino do hiperparâmetro de decaimento, que pode ser sensível ao conjunto de dados e à arquitetura do modelo (MEHMOOD; AHMAD; WHANGBO, 2023).

- Adam (KINGMA; BA, 2014): abreviação de *Adaptive Moment Estimation*, é um otimizador amplamente utilizado devido à sua eficiência e estabilidade no treinamento de Redes Neurais Profundas. Ele combina as vantagens do RMSprop e do SGD, ajustando dinamicamente a taxa de aprendizado com base em estimativas de primeira e segunda ordem dos gradientes. Isso resulta em uma convergência rápida e desempenho consistente em problemas não estacionários (REDDI; KALE; KUMAR, 2019).

- AdaMax (KINGMA; BA, 2014): é uma variante do otimizador Adam que utiliza normas infinitas para atualizar os pesos, proporcionando maior estabilidade ao lidar com gradientes amplamente distribuídos. Ele mantém os benefícios de adaptação dinâmica da taxa

de aprendizado do Adam, mas apresenta uma convergência mais robusta em cenários com gradientes irregulares, especialmente em tarefas com hiperparâmetros sensíveis (DOGO et al., 2018). Comparado ao Adam, o AdaMax apresenta melhor desempenho em redes neurais profundas que exigem maior regularização e estabilidade (ARIFF; ISMAIL, 2023).

2.6 Redes Neurais Convolucionais

As Redes Neurais Convolucionais (*Convolutional Neural Network*, em inglês – CNN), representam uma variação das MLPs inspirada na maneira biológica de processamento de informações. Embora as CNNs sejam utilizadas desde a década de 1980, avanços significativos em áreas como Capacidade Computacional e Disponibilidade de Dados permitiram que, hoje, essas redes superem habilidades humanas em determinadas tarefas complexas de Visão Computacional (GÉRON, 2019). Além disso, destacam-se em campos como Reconhecimento de Voz (*Voice Recognition*, em inglês – VR), Processamento de Linguagem Natural (*Natural Language Processing – NLP*(GERING; CIARELLI; SALLES, 2019) e em aplicações que envolvem dados multidimensionais, como, por exemplo, vídeos (FOGAÇA et al., 2022).

Nas etapas iniciais de uma CNN, encontram-se as camadas de convolução e *pooling*, cuja função principal é extrair e abstrair as características das imagens. Posteriormente, uma camada completamente conectada, baseada em uma MLP, assume o papel de realizar a classificação. A Figura 2.8 apresenta um exemplo prático de uma CNN e a disposição de suas camadas.

O nome *"Rede Neural Convolucional"* indica que a rede emprega uma operação matemática chamada **convolução**, que é um tipo especializado de operação linear. CNNs são simplesmente redes neurais que utilizam, em pelo menos uma de suas camadas, a operação de convolução no lugar da matriz de multiplicação convencional (GOODFELLOW; BENGIO; COURVILLE, 2016).

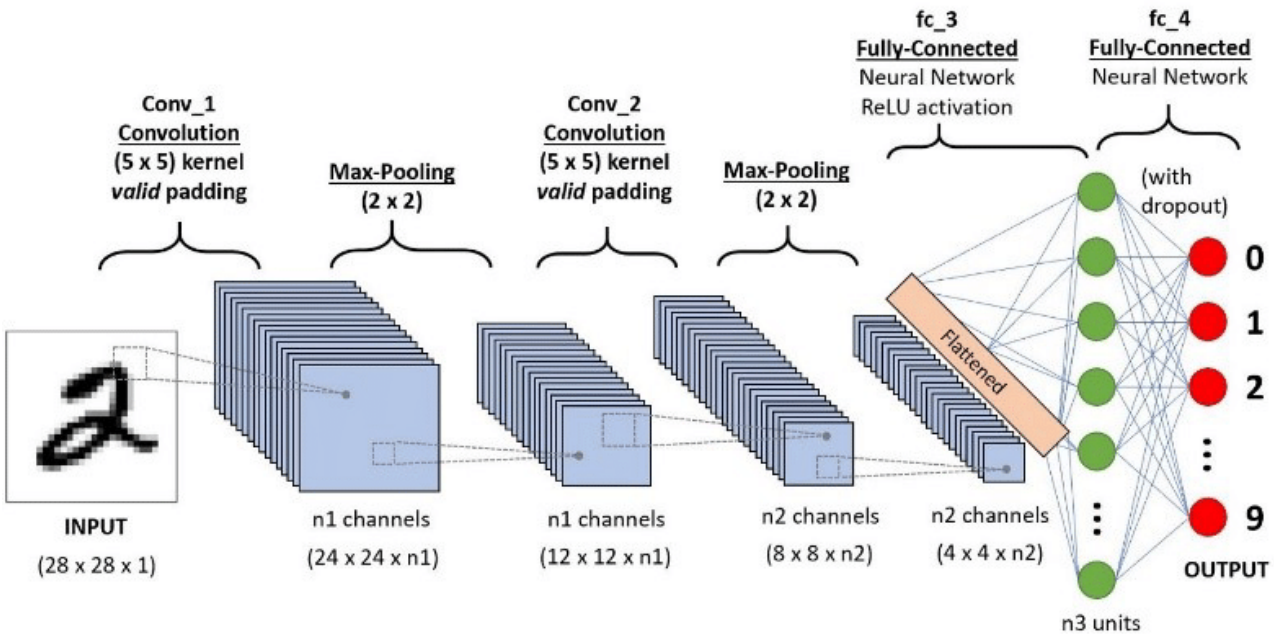

Figura 2.8: Exemplo de CNN (SHAH et al., 2020).

2.6.1 Camadas de Convolução

A operação de convolução é considerada a mais significativa dentro de uma CNN. Essa etapa aplica a operação de convolução, definida pela Equação 2.11, que expressa a quantidade de sobreposição de uma função g à medida que ela é deslocada sobre outra função f. Portanto, pode-se generalizar que a convolução *combina* uma função com outra (WISSTEIN, 2024). As camadas da CNN que realizam essa operação são referidas como **camadas convolucionais** ou **camadas de convolução**.

$$f * g = \int f(\tau)g(t-\tau)d\tau \tag{2.11}$$

Na terminologia de CNN, a função f da Equação 2.11 pode ser referida como *input* (entrada), e a função g como *kernel* (núcleo) ou filtro, e o resultado (*output*) da convolução como *feature map* (mapa de características). Em aplicações de ML, usualmente a entrada é um vetor multidimensional de dados, enquanto o núcleo é usualmente um vetor multidimensional de parâmetros, parâmetros esses que são adaptados pelo algoritmo de aprendizagem (GOODFELLOW; BENGIO; COURVILLE, 2016). A Figura 2.9 exemplifica a operação de convolução para

um filtro (3×3).

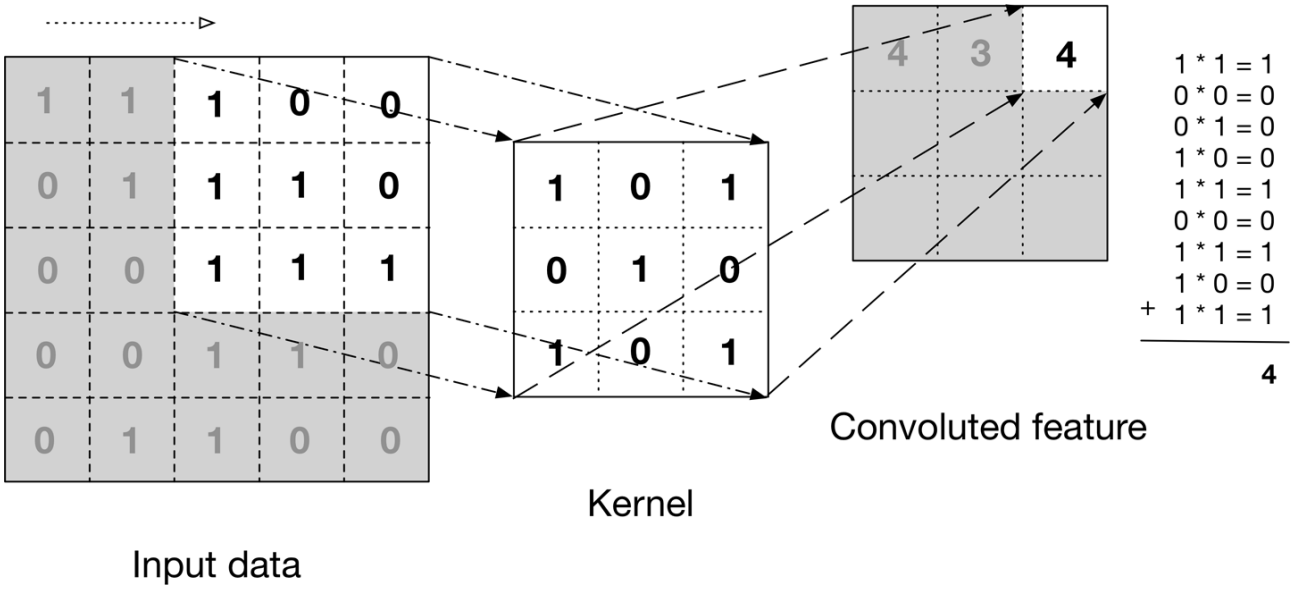

Figura 2.9: Convolução de uma entrada (5×5) com a aplicação de um *kernel* (3×3) (MANAV, 2024).

Quando uma imagem é fornecida como entrada para uma CNN, os dados dessa imagem são representados por *pixels*. Esses *pixels*, conforme o esquema de cores RGB, possuem três componentes inteiras distintas. Com isso, uma imagem pode ser representada puramente como uma matriz tridimensional que inclui comprimento, largura e três canais de cor, conforme ilustra a Figura 2.10. Dessa forma, as camadas de convolução da CNN são capazes de operar a partir das representações numéricas dos *pixels* da imagem.

2.6.2 Camadas de Pooling

Geralmente, tem-se como camadas subsequentes das CNNs as camadas de *pooling*, cuja função é reduzir a dimensão de entrada provida pela camada anterior. Usualmente, utiliza-se dois tipos principais de camadas deste tipo, a de *Max Pooling* e a de *Mean Pooling* (KHAN et al., 2018). A camada de *Max Pooling* produz como saída o maior valor fornecido dentre o conjunto de valores fornecidos como entrada. A de *Mean Pooling*, por sua vez, retorna como resultado a média aritmética dos valores de entrada. A Figura 2.11 exemplifica, de forma simplificada, as operações realizadas para esses dois tipos de camadas de *pooling*.

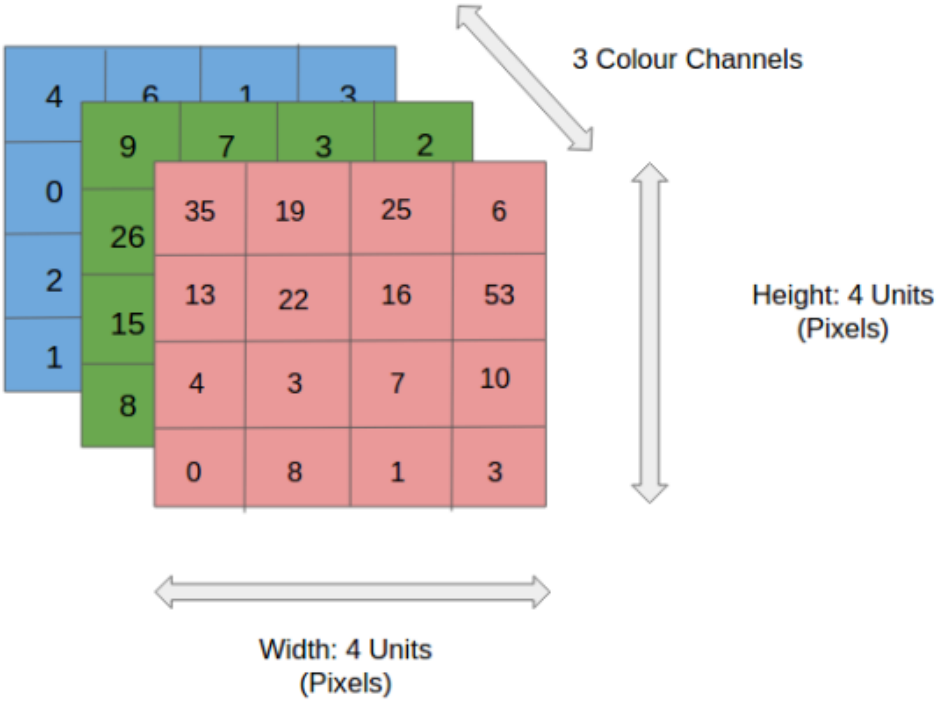

Figura 2.10: Imagem representada a partir de uma matriz tridimensional (MANAV, 2024).

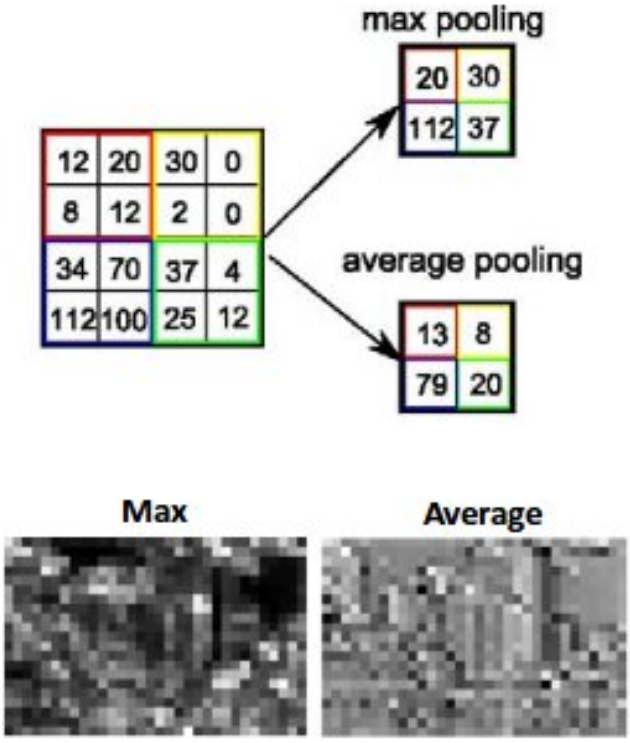

Figura 2.11: *Max Pooling* e *Mean Pooling* (RAHMAN, 2018).

2.6.3 Camadas Totalmente Conectadas

As camadas finais de uma CNN são totalmente conectadas (*fully connected*) e têm como função principal gerar a probabilidade associada à classificação da imagem. Essa probabilidade é

obtida através de uma função de ativação, sendo a *softmax* a mais utilizada para esse propósito (WOOD, 2018). Nessa estrutura, todos os neurônios de uma camada estão interligados aos neurônios da camada seguinte, de forma semelhante ao funcionamento de uma MLP tradicional. O número de neurônios presentes geralmente corresponde à quantidade de classes a serem reconhecidas. Adicionalmente, essas camadas podem incluir a técnica de *dropout*, que desativa aleatoriamente alguns neurônios durante o treinamento, reduzindo o risco de sobreajuste (*overfit*) e melhorando a generalização do modelo (KRIZHEVSKY; SUTSKEVER; HINTON, 2017).

2.7 Arquiteturas de CNN

As arquiteturas padrão de CNN seguem uma organização em que as camadas são estruturadas da seguinte maneira (GÉRON, 2019):

- Camadas convolucionais, acompanhadas por uma função de ativação;
- Uma camada de *pooling*;
- Mais camadas convolucionais, também seguidas por funções de ativação;
- Outra camada de *pooling*.

Esse padrão se repete até alcançar as camadas totalmente conectadas. À medida que a imagem é processada por essas camadas, suas dimensões diminuem gradualmente, enquanto suas representações se tornam mais profundas e específicas.

Com o progresso contínuo no campo das redes neurais, diversas variações da arquitetura clássica de CNN foram desenvolvidas ao longo dos anos. Esse avanço pode ser observado em grandes competições como o *Desafio de Reconhecimento Visual em Grande Escala do ImageNet* (ILSVRC), que demonstra uma considerável redução na taxa de erro para classificação de imagens: de 26% para menos de 3% em apenas seis anos (RUSSAKOVSKY et al., 2015).

Baseado no extenso conjunto de dados ImageNet, que inclui mais de 14 milhões de imagens rotuladas em mais de 20.000 categorias, o *benchmark* (CODE, 2024) compara o desempenho de

diversos tipos de modelos na tarefa de classificação de imagens do ImageNet, e é constantemente atualizado de modo a incluir novos modelos referentes ao estado da arte atual.

Nas Seções 2.7.1 a 2.7.3 são descritas três arquiteturas de CNN que se destacam no *benchmark* (CODE, 2024), pois na tarefa de classificação de imagens no ImageNet obtiveram ótimos resultados: Inception (SZEGEDY et al., 2016a), EfficientNet (TAN; LE, 2019) e ConvNext (LIU et al., 2022).

2.7.1 Inception

A arquitetura Inception, também conhecida como GoogLeNet, foi apresentada durante a competição ILSVRC de 2014, onde conquistou o primeiro lugar com uma taxa de erro de apenas 6,67%. Essa rede possui 22 camadas e utiliza cerca de quatro milhões de parâmetros, representando um avanço significativo em comparação aos 60 milhões de parâmetros da AlexNet (SZEGEDY et al., 2016b). A Inception é formada por blocos convolucionais denominados *Inception Modules*, organizados de modo que suas saídas variem estatisticamente, permitindo que as camadas superiores capturem características mais abstratas. Além disso, ela incorpora filtros de diferentes tamanhos, como $1x1$, $3x3$ e $5x5$, conforme ilustrado pela Figura 2.12 (SZEGEDY et al., 2016b).

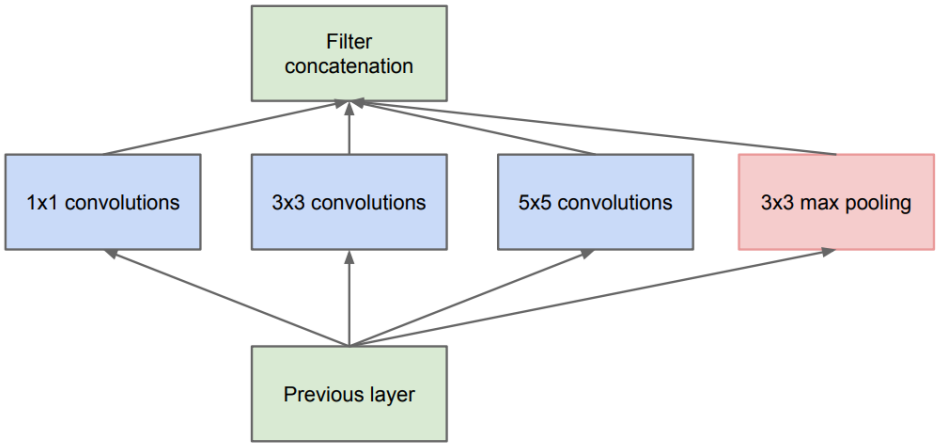

Figura 2.12: Composição de um *Inception Module* (SZEGEDY et al., 2016b).

2.7.2 EfficientNet

A arquitetura EfficientNet foi a vencedora da competição ILSVRC de 2019, atingindo uma taxa de erro de apenas 2,9%. Foi considerada o estado da arte da época. Tal arquitetura demonstrou precisão superior com uma redução significativa no número de parâmetros e no consumo computacional em comparação a arquiteturas como ResNet e Inception (Seção 2.7.1, alcançando uma precisão superior com até 10 vezes menos parâmetros. Esse avanço consolidou seu papel em áreas como Visão Computacional e Aprendizado Profundo (TAN; LE, 2019).

A principal característica da EfficientNet é o seu método de escalonamento composto (*compound scaling*) projetado para balancear e otimizar simultaneamente a profundidade, largura e resolução da rede neural, diferentemente de modelos tradicionais que escalavam esses fatores de forma isolada (TAN; LE, 2019). A Figura 2.13 ilustra a sequência de camadas do modelo EfficientNet-B0, a versão mais leve da arquitetura.

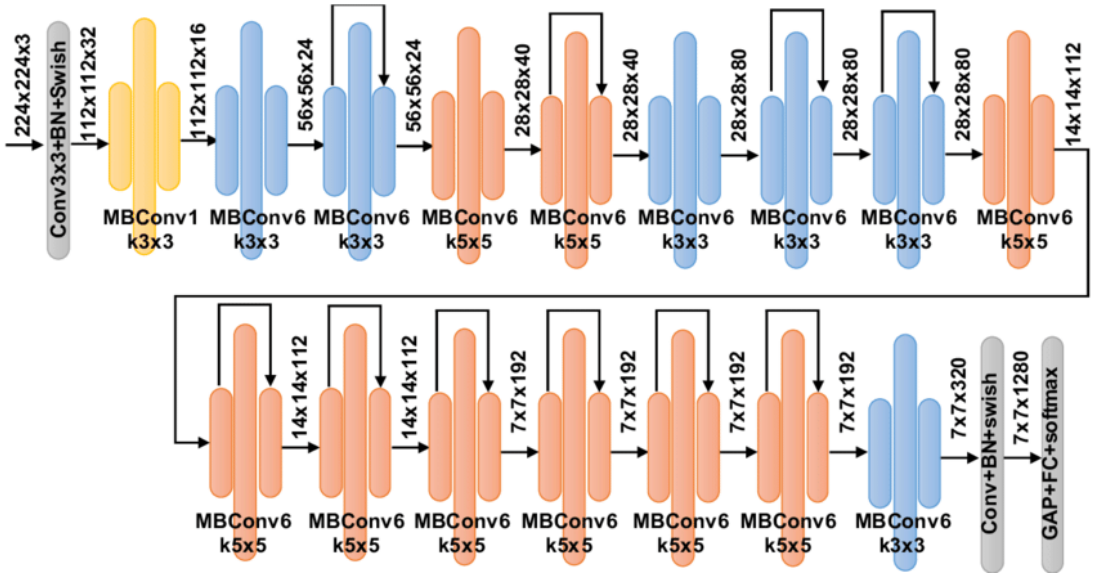

Figura 2.13: Sequência de camadas do modelo EfficientNet-B0 (ALHICHRI et al., 2021).

2.7.3 ConvNeXt

A recente arquitetura ConvNeXt foi desenvolvida com o objetivo de modernizar as CNNs à luz dos avanços em arquiteturas baseadas em *Transformers*, como, por exemplo, o *Vision Transformer* (ViT) (veja Seção 2.8) (ALEXEY, 2020). Inspirada por inovações trazidas por essas

novas abordagens, a ConvNeXt ajusta e refina elementos tradicionais das CNNs, para alcançar um desempenho competitivo com os *Transformers*, mantendo a eficiência computacional inerente às CNNs. A proposta introduziu melhorias incrementais, como, por exemplo, convoluções em profundidade, *activation scaling* e estratégias de *residual connection*. Essas melhorias aumentam a flexibilidade e o poder expressivo do modelo (LIU et al., 2022).

Em *benchmarks* relevantes como o ImageNet, a ConvNeXt alcançou resultados comparáveis aos dos *Transformers* mais avançados, com menores custos computacionais, o que a torna uma opção atraente para sistemas com recursos limitados. Além disso, mostrou versatilidade em tarefas de detecção de objetos e segmentação semântica, consolidando-se como uma alternativa eficiente para arquiteturas baseadas em *Transformers*. A Figura 2.14 apresenta a sequência de módulos da arquitetura ConvNeXt.

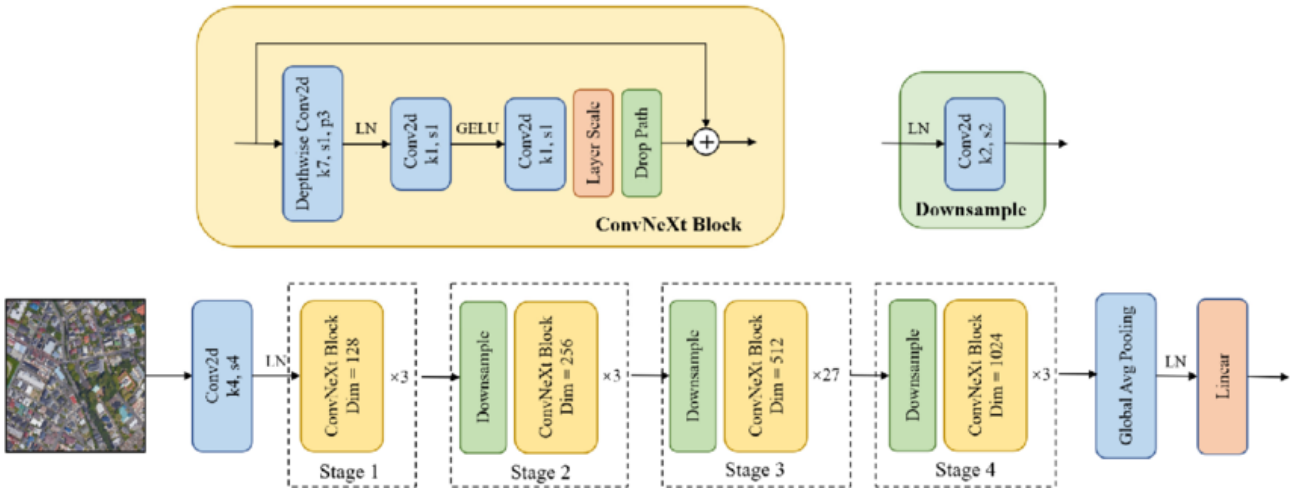

Figura 2.14: Sequência de blocos da arquitetura ConvNeXt (CHEN et al., 2023).

2.8 *Transformers*

Os *Transformers* (transformadores, em português) são uma classe de modelos de aprendizado profundo projetados para processar dados sequenciais. Introduzidos por (VASWANI, 2017), os *Transformers* revolucionaram áreas como Processamento de Linguagem Natural (do inglês, *Natural Language Processing* – NLP) e Visão Computacional, tornando-se a base para modelos como BERT (KENTON; TOUTANOVA, 2019), GPT (RADFORD, 2018) e *Vision Transformer*

(ViT) (ALEXEY, 2020). A principal inovação dos *Transformers* é o mecanismo de *attention*, que permite que o modelo aprenda dependências de longo alcance em sequências de dados sem usar convoluções ou recursividade. Nas Seções 2.8.1 a 2.8.4, serão abordados os principais conceitos e características que compõem a arquitetura dos *Transformers*, que é representada pela Figura 2.15.

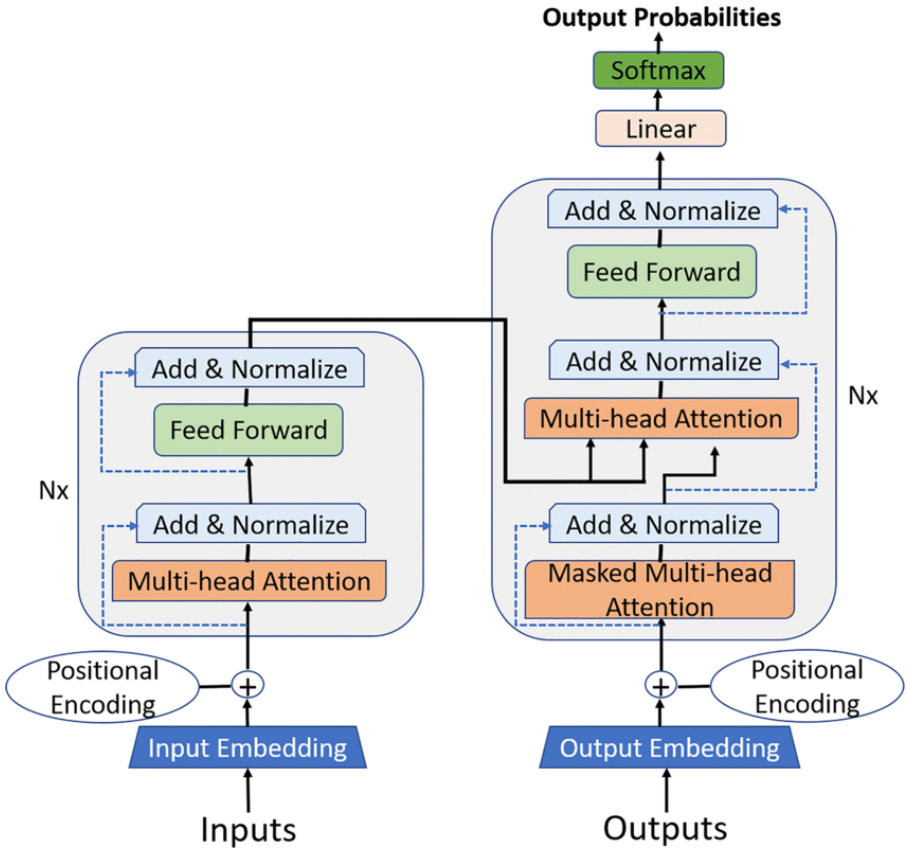

Figura 2.15: Arquitetura do *Transformer* (NASSIRI; AKHLOUFI, 2023).

.

2.8.1 Embeddings

Embeddings são representações densas e contínuas de dados em um espaço de menor dimensão, projetadas para capturar características semânticas e relações contextuais entre os elementos representados (BENGIO; DUCHARME; VINCENT, 2000). Isso permite que modelos *Transformers* processem a entrada de forma matemática e aprendam relações semânticas ou sintáticas entre os elementos da sequência. Por isso, o *embedding* é crucial para transformar entradas

textuais em representações utilizáveis por redes neurais (VASWANI, 2017).

Um exemplo clássico de *embeddings* é o Word2Vec (MIKOLOV, 2013), que posiciona palavras em um espaço vetorial de forma que palavras com significados semelhantes fiquem próximas. Embora o *Transformer* não utilize diretamente *embeddings* pré-treinados, ele emprega conceitos semelhantes para iniciar os vetores de entrada.

Inicialmente, cada *token* da entrada é transformado em um vetor. Essa transformação é feita por meio de uma camada de *embedding* que é treinada juntamente com o modelo. Essa camada aprende a representar os *tokens* de maneira otimizada para a tarefa específica, como tradução ou geração de texto (VASWANI, 2017).

Como o *Transformer* não é sequencial e, portanto, não possui mecanismos intrínsecos para capturar a ordem dos *tokens*, a informação posicional é adicionada aos *embeddings*. Isso é feito somando-se os vetores de *embeddings* com codificações posicionais (*positional encodings*) calculadas utilizando funções trigonométricas. Essa combinação permite que o modelo diferencie a posição dos *tokens* na sequência, capturando dependências de ordem em tarefas como tradução e resumo de texto (VASWANI, 2017).

2.8.2 Codificador (*Encoder*)

Na arquitetura *Transformer*, o codificador (*encoder*, em inglês) é projetado para processar a entrada – por exemplo, uma sequência de texto ou *patches* de uma imagem – e transformá-la em uma representação contextual rica. Essa representação é posteriormente utilizada pelo decodificador (*decoder*, em inglês) em tarefas de geração de saída, como tradução de idiomas ou resumo (VASWANI, 2017).

O codificador consiste em uma pilha de N camadas idênticas, e cada camada é composta por dois subcomponentes principais: o Mecanismo de Atenção *Multi-head* (*Multi-head Self-Attention*, em inglês), e a Rede *Feedforward* Posicional (*Position-wise Feedforward Network*, em inglês). Cada uma dessas partes é envolvida por camadas de normalização (*Add & Normalize*), e as conexões residuais ajudam a estabilizar o treinamento (KENTON; TOUTANOVA, 2019).

O objetivo do mecanismo de atenção (*Multi-head Self-Attention*) é calcular como cada *token*

(ou parte da entrada) está relacionado a todos os outros *tokens* da sequência. Isso é feito calculando três matrizes aprendidas: Consulta (*Query*, Q), Chave (*Key*, K) e Valor (*Value*, V). Dessa forma, a atenção é calculada conforme a Equação 2.12, onde $\sqrt{d_k}$ é o fator de escala que estabiliza os cálculos de atenção (VASWANI, 2017).

$$Attention(Q, K, V) = softmax\left(\frac{QK^T}{\sqrt{d_k}}\right)V \tag{2.12}$$

Após o mecanismo de atenção, os dados passam pela Rede *Feedforward* Posicional (*Position-wise Feedforward Network*, em inglês), que consiste em duas camadas totalmente conectadas com uma função de ativação ReLU (Seção 2.4). Essa operação é aplicada de forma independente a cada posição na sequência, permitindo transformações não lineares nas representações dos *tokens* (VASWANI, 2017).

2.8.3 Decodificador (*Decoder*)

O decodificador é projetado para gerar as sequências de saída, como, por exemplo, traduções de texto, legendas de imagem ou respostas em um *chatbot*. Ele trabalha em sinergia com o codificador, utilizando os *embeddings* gerados a partir da sequência de entrada para prever iterativamente os próximos *tokens* na saída. O decodificador combina **Atenção Própria Mascarada** (*Masked Self-Attention*, em inglês – MSA) com **Atenção Cruzada** (*Cross-Attention*, em inglês – CA) para integrar informações da entrada e gerar saídas coerentes e contextuais (VASWANI, 2017).

Cada decodificador consiste em uma pilha de camadas idênticas, e possui três subcomponentes principais em cada camada: MSA, CA e uma Rede *Feedforward*. O mecanismo MSA calcula dependências entre os *tokens* já gerados na sequência de saída, garantindo que o modelo seja autoregressivo, ou seja, que cada previsão dependa apenas dos *tokens* anteriores e não daqueles que ainda serão gerados. Essa restrição é implementada por uma máscara causal, que impede que o modelo "olhe para o futuro". A CA, por sua vez, conecta os *tokens* gerados pelo decodificador com as representações contextuais da entrada processada pelo codificador, permitindo que o modelo relacione diretamente a sequência de entrada com a saída (VASWANI, 2017).

Além dos mecanismos de atenção, o decodificador inclui Redes *Feedforward* Posicionais (*Position-wise Feedforward Network*, em inglês – PwFN) em cada camada. Assim como no codificador, essas redes são formadas por camadas densas totalmente conectadas e introduzem complexidade adicional às interações não lineares entre os *tokens*, refinando as representações. Além disso, camadas de normalização e conexões residuais encapsulam cada subcomponente, o que melhora a estabilidade do treinamento e facilita a propagação do gradiente. Após passar por todas as camadas do decodificador, a sequência resultante é transformada em distribuições de probabilidade sobre o vocabulário por uma camada linear seguida de uma função *softmax*, permitindo que o modelo escolha o próximo *token* na sequência (VASWANI, 2017).

A interação entre MSA e CA torna o decodificador extremamente eficaz em tarefas de geração de sequência condicionada. Por exemplo, em tradução automática, ele utiliza a entrada processada pelo codificador para gerar a sentença traduzida palavra por palavra. Sua estrutura modular e flexível também permite aplicações em diversas tarefas, como geração de texto e legendagem automática (VASWANI, 2017).

2.8.4 *Vision Transformers* (ViTs)

Os Vision Transformers (ViTs) representam uma adaptação inovadora da arquitetura *Transformer* para tarefas de visão computacional, como, por exemplo, classificação de imagens, detecção de objetos e segmentação. Introduzidos por (ALEXEY, 2020), os ViTs desafiaram a predominância das redes convolucionais (CNNs) em visão computacional, provando que mecanismos de atenção baseados em *Transformers* podem competir ou superar as CNNs quando treinados em grandes conjuntos de dados.

Os ViTs seguem a arquitetura básica dos *Transformers* descrita por (VASWANI, 2017), mas adaptam o fluxo de dados para lidar com entradas bidimensionais (imagens), em vez de sequências unidimensionais como texto. A principal inovação está na transformação das imagens em uma representação compatível com o *Transformer*, feita por meio de componentes chamados *Patch Embeddings* (ALEXEY, 2020).

A primeira etapa do *Patch Embedding* consiste em dividir a imagem em pequenos blocos

chamados *patches*. Cada *patch* é uma seção da imagem de tamanho fixo, como 16×16 *pixels*, representando uma "palavra visual". Assim como os *tokens* em NLP, esses *patches* formam uma sequência linear que pode ser processada pelo modelo *Transformer*. Por exemplo, uma imagem 224×224 pode ser dividida em $14 \times 14 = 196$ *patches* de 16×16 *pixels*. Cada *patch* é achatado em um vetor unidimensional de dimensão $16 \times 16 \times C$, onde C é o número de canais da imagem (usualmente 3, para imagens RGB).

Após a divisão em *patches*, cada *patch* é transformado em um vetor de dimensão fixa usando uma camada linear aprendida, conhecida como camada de *patch embedding*. Essa camada mapeia os *patches* do espaço original para o espaço da dimensão usada pelo *Transformer*.

Assim como no *Transformer clássico*, os ViTs usam codificação posicional para garantir que a posição relativa dos *patches* na imagem seja levada em consideração. Isso se dá pelo fato de que os *patches* são processados como uma sequência, e a ausência de convoluções requer uma indicação explícita de ordem ou localização espacial (ALEXEY, 2020). As etapas para que os dados de entrada cheguem até o codificador do *Transformer* são apresentadas na Figura 2.16

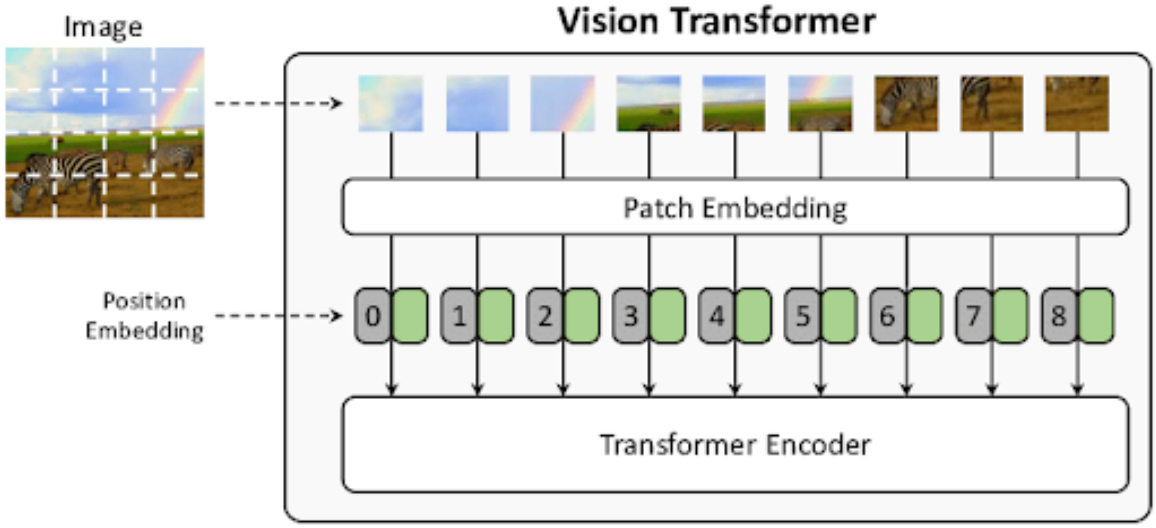

Figura 2.16: Atuação do *Patch Embedding* no *Vision Transformer* (STEFANINI et al., 2022).

Por fim, após passar pelas camadas do *Transformer*, a saída do *token* que contém a representação final da imagem é passada por uma camada linear ou um classificador *softmax* (Seção 2.5) para produzir as previsões finais do modelo (ALEXEY, 2020).

2.9 Arquiteturas de ViT

Conforme descrito na Seção 2.7, o *benchmark* (CODE, 2024) compara o desempenho de diferentes tipos de modelos para a tarefa de classificação de imagens do ImageNet. Nos últimos anos, os modelos de ViT chegaram a ultrapassar o desempenho das CNNs no *benchmark*, de modo que, até o presente momento, os dez primeiros modelos do ranking atual são modelos de ViT, representando assim o estado da arte atual no desafio do ImageNet de classificação de imagens.

Nesse contexto, as seções 2.9.1 até 2.9.3 apresentam os modelos da arquitetura *Vision Transformer* que foram selecionados, pois apresentaram desempenho excepcional no desafio do ImageNet de classificação de imagens.

2.9.1 ViT

A ViT (*Vision Transformer*) (ALEXEY, 2020) é uma arquitetura pioneira que adapta a estrutura dos transformadores, amplamente utilizados em Processamento de Linguagem Natural (PLN), para o domínio da visão computacional. Sua principal inovação é tratar imagens como sequências de *patches*, semelhante ao tratamento de palavras em um texto. Cada imagem é dividida em pequenos blocos fixos (*patches*), que são linearizados e, em seguida, processados por um mecanismo de autoatenção para capturar as relações globais entre eles, conforme ilustrado na Figura 2.17. Essa abordagem permite que o modelo aprenda representações visuais de maneira direta, sem depender de convoluções, o que facilita sua escalabilidade em tarefas complexas, como, por exemplo, classificação de imagens em conjuntos de dados grandes como o ImageNet (ALEXEY, 2020).

A simplicidade e eficiência da ViT se destacam, especialmente, quando combinada com grandes volumes de dados de treinamento. Recentemente, a arquitetura vem se tornando base para a construção de outros novos modelos de *Vision Transformer* adaptados. Resultados experimentais registrados por (ALEXEY, 2020) mostram que, com um treinamento adequado e dados suficientes, a ViT supera arquiteturas convolucionais tradicionais em precisão, consolidando-se como uma alternativa revolucionária no design de modelos de visão computacional.

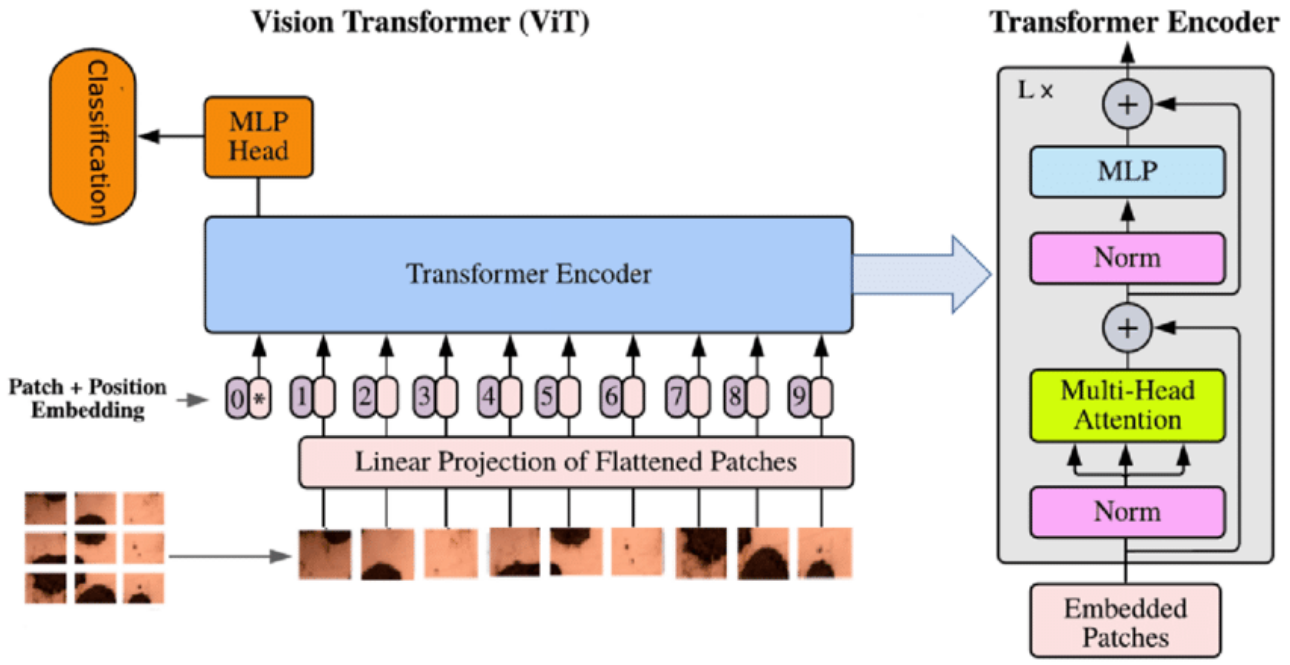

Figura 2.17: Modelo de arquitetura do ViT (ALRFOU; KORDIJAZI; ZHAO, 2022).

2.9.2 MaxViT

O MaxViT (abreviação de *Maximum Vision Transformer*) (TU et al., 2022) é um modelo inovador que combina transformadores e convoluções para processar imagens de maneira eficiente e escalável. Sua arquitetura hierárquica utiliza blocos *Transformer* para capturar relações globais e locais em imagens, equilibrando precisão e eficiência computacional. Uma das suas inovações é o uso de atenção axial (HO et al., 2019), que reduz a complexidade computacional ao dividir a atenção em dimensões horizontais e verticais, enquanto preserva informações detalhadas por meio de convoluções locais. A Figura 2.18 apresenta os blocos que compõem o MaxViT.

O modelo demonstrou desempenho superior em *benchmarks*, como ImageNet, comparado a modelos populares como *Swin Transformer* (LIU et al., 2021) e *EfficientNet*. Além de sua precisão, destaca-se pela escalabilidade e eficiência em dispositivos modernos, facilitando sua aplicação em diagnósticos médicos, veículos autônomos e outras áreas que exigem análise visual detalhada. O projeto modular do MaxViT também permite fácil adaptação para diferentes tamanhos de entrada e demandas computacionais, consolidando sua relevância na visão computacional contemporânea (TU et al., 2022).

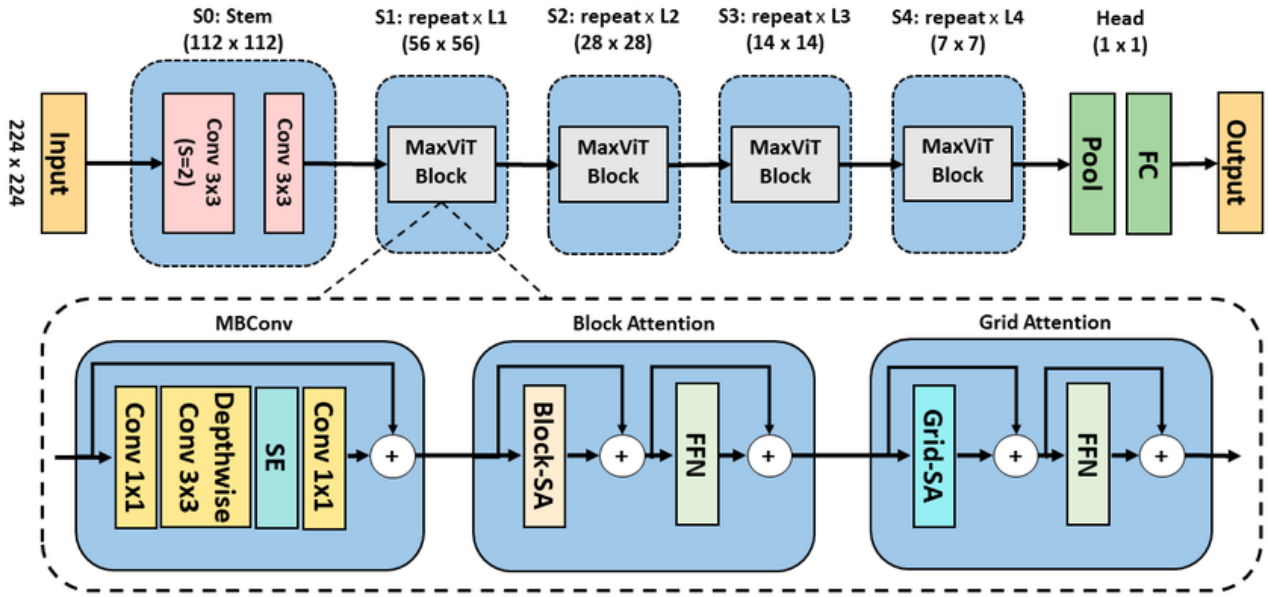

Figura 2.18: Arquitetura modelo do MaxViT (ONG et al., 2024).

2.9.3 DaViT

A arquitetura DaViT (*Dual Attention Vision Transformer*) (DING et al., 2022) é uma abordagem inovadora no campo de modelos de visão computacional que une mecanismos de atenção local e global. A arquitetura utiliza uma combinação estratégica de transformadores e convoluções para lidar com a hierarquia espacial das imagens de maneira eficaz. Em especial, o DaViT emprega mecanismos de atenção dual, que equilibram a captura de relações locais e globais, otimizando o custo computacional sem comprometer o desempenho. Isso é obtido por meio de uma arquitetura modular, na qual cada módulo é projetado para processar características específicas, como detalhes finos em resoluções mais baixas e informações contextuais amplas em resoluções mais altas (DING et al., 2022), conforme ilustra a Figura 2.19.

A implementação do DaViT também se destaca por sua adaptabilidade a diferentes tamanhos de modelos e aplicações, desde tarefas de classificação até segmentação de imagens. A integração de convoluções e transformadores contribui para superar limitações de escalabilidade e eficiência encontradas em abordagens tradicionais. *Benchmarks* relevantes como ImageNet mostram que os modelos da arquitetura DaViT alcançam desempenho competitivo em *benchmarks* como ImageNet, destacando-se como uma alternativa poderosa para modelos puramente

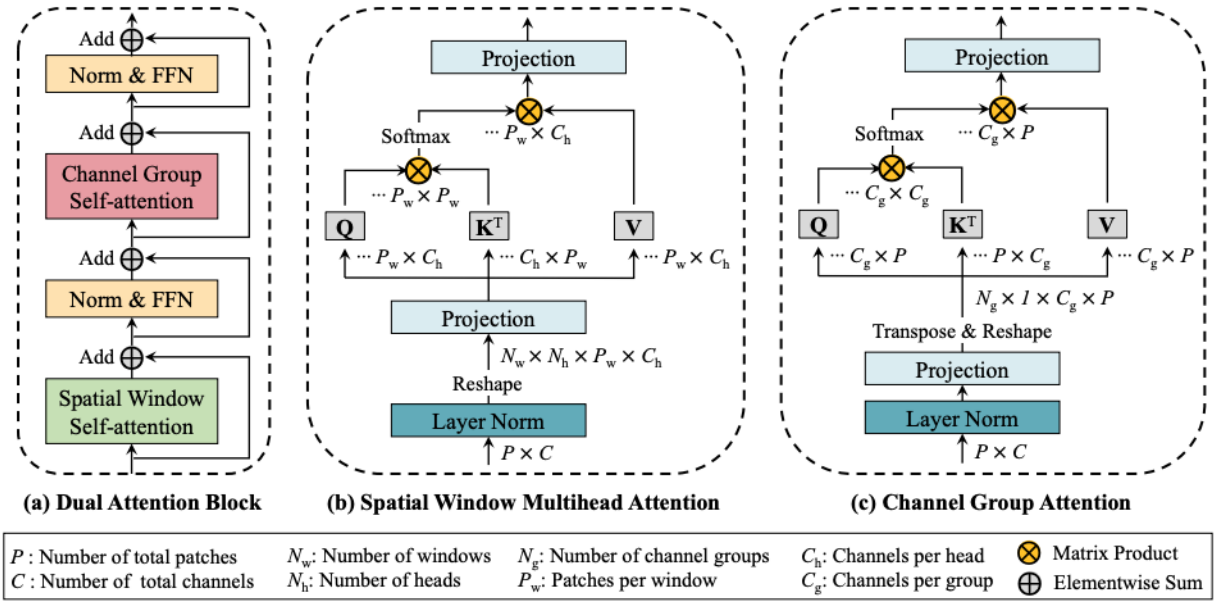

Figura 2.19: Módulos componentes do DaViT (DING et al., 2022).

baseados em atenção ou convoluções. Essa arquitetura reflete um avanço significativo no design de modelos híbridos, atendendo às demandas crescentes de eficiência computacional em visão computacional moderna.

2.10 Definições de Métodos e Estratégias de Implementação

2.10.1 Modelos Pré-Treinados

Treinar uma rede neural do zero com um conjunto de dados tende a ser desafiador, pois a insuficiência de exemplos compromete o desempenho do modelo. Para superar essa limitação, geralmente são empregados modelos pré-treinados, de modo a aproveitar o conhecimento prévio para extrair características relevantes das imagens. Além disso, essa abordagem favorece uma melhor generalização, já que o modelo pré-treinado foi desenvolvido com base em um conjunto de dados amplo e diversificado.

A ampla base de imagens fornecida pelo ImageNet tornou viável o desenvolvimento e treinamento de redes neurais profundas mais sofisticadas, capazes de apresentar resultados significativos. Entretanto, o processo de treinamento dessas redes exige um investimento considerável

de tempo, mesmo quando se utiliza *hardware* de alto desempenho. Nesse contexto, *benchmarks* e competições como o ImageNet têm grande relevância acadêmica, pois permitem que diversos grupos de pesquisa compartilhem modelos avançados, treinados durante extensos períodos com o apoio de infraestrutura tecnológica robusta e grandes conjuntos de dados.

Em nosso trabalho, são empregados seis modelos pré-treinados a partir da base de dados do ImageNet. Desses, três são CNNs e outros três ViTs, todos de código aberto e amplamente acessíveis à comunidade. As CNNs selecionadas foram: Inception-v3 (SZEGEDY et al., 2016b), EfficientNet-V2 (TAN; LE, 2021) e ConvNeXt-L (LIU et al., 2022). Os *Vision Transformers* selecionados foram: ViT-L/16 (ALEXEY, 2020), MaxViT-T (TU et al., 2022) e DaViT-B (DING et al., 2022). As especificidades de cada modelo foram detalhadas nas seções 2.7 e 2.9.

2.10.2 Transferência de Aprendizado – *Transfer Learning*

Essa técnica consiste em utilizar padrões previamente aprendidos na solução de outros problemas para reduzir o tempo de treinamento e os custos relacionados ao uso de hardware (PAN; YANG, 2010).

Conforme cita (MILANI et al., 2023), há três estratégias principais para a utilização da transferência de aprendizado:

- Uso do modelo pré-treinado como extrator de características: essa abordagem consiste em empregar um modelo pré-treinado, geralmente desenvolvido com base em um conjunto de dados extenso e genérico, como, por exemplo, o ImageNet, para obter as características das novas imagens. Posteriormente, essas características extraídas são utilizadas para treinar um modelo adaptado a um conjunto de dados menor e mais específico. Nesse processo, a última camada do modelo é substituída por uma camada densa com o número de unidades correspondente ao total de classes do novo conjunto de dados. Destaca-se que nessa estratégia há uma redução no risco de *overfitting*, pois apenas os pesos da camada de classificação precisam ser ajustados.

- Ajuste fino do modelo pré-treinado: consiste em adaptar um modelo pré-treinado para

atender às necessidades de um novo conjunto de dados. O objetivo é aproveitar o conhecimento adquirido pelo modelo em um conjunto de dados genérico e refiná-lo para se adequar a um contexto mais específico. Para isso, algumas camadas do modelo original são mantidas congeladas, enquanto que outras são ajustadas em conjunto com as novas camadas adicionadas ao modelo.

- Treinamento de rede com iniciação por transferência: essa abordagem começa com o treinamento de uma nova rede, cuja arquitetura geralmente se assemelha a de um modelo pré-treinado, mas com pesos inicialmente atribuídos de forma aleatória. Posteriormente, os pesos do modelo pré-treinado são carregados para iniciar a rede. A partir desse ponto, o treinamento é conduzido utilizando o conjunto de dados específico.

2.10.3 Validação Cruzada K-*Fold*

A validação cruzada K-*Fold* consiste em dividir o conjunto completo de dados em k partes, chamadas de *folds*. Esses *folds* são usados alternadamente para treinamento e validação, permitindo que cada parte do conjunto de dados seja reservada como conjunto de teste em pelo menos uma iteração. Ao término do processamento, as k medições de desempenho são combinadas, geralmente calculando-se a média e o desvio padrão, para obter uma estimativa consolidada do desempenho final (SILVA, 2021). A Figura 2.20 ilustra a atuação do K-*Fold* com $k = 5$.

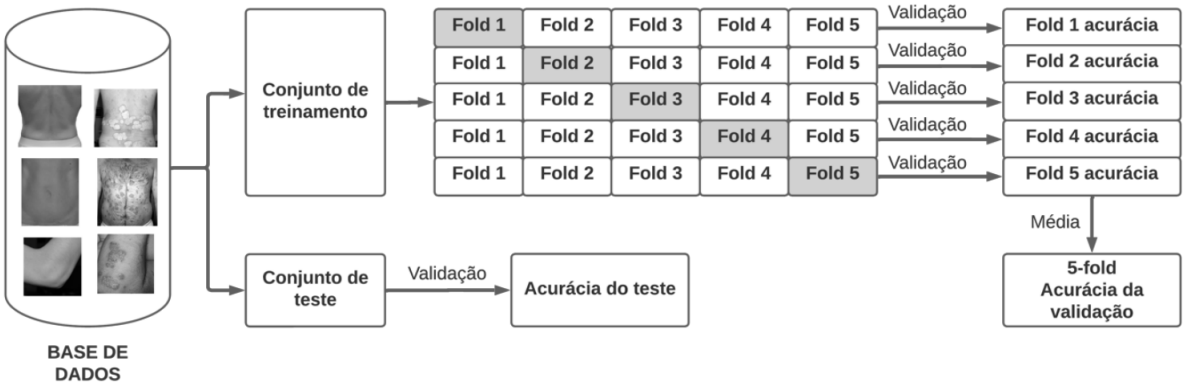

Figura 2.20: Validação cruzada K-*Fold* com k sendo cinco (MILANI et al., 2023).

2.10.4 Algoritmo *dHash*

Um importante fator que deve ser considerado na construção de um conjunto de dados extraídos de múltiplas fontes distintas é a possibilidade de existirem imagens duplicadas no conjunto. Isso é porque esse cenário é prejudicial para o treinamento dos modelos de aprendizagem supervisionada, pois torna o conjunto enviesado. Dessa forma, durante a fase de treino, a rede neural pode obter mais oportunidades de aprender padrões específicos das duplicatas do que para o restante das imagens. Isso prejudica a capacidade do modelo de generalizar para uma nova imagem que não pertence ao conjunto de dados de treino (MANJUSHA, 2022).

A implementação do algoritmo *Differencial Hash* (*dHash*) é um método preciso e eficiente para a detecção de imagens duplicadas no conjunto de dados (MANJUSHA, 2022). O algoritmo gera uma representação numérica (*hash*) para cada imagem, de forma que imagens que geram o mesmo *hash* são consideradas duplicatas.

Para calcular o *hash* de cada imagem, inicialmente o algoritmo redimensiona a imagem para um tamanho $(n+1) \times n$, garantindo que haverá 01 (um) pixel a mais de largura para armazenar um campo de diferença, utilizado para gerar a matriz de campo de diferença. Por questões de desempenho e convenção, o valor escolhido para a execução do algoritmo sobre o conjunto de dados foi $n = 8$, redimensionando assim cada imagem para a dimensão 9×8. Nessa etapa de redimensionamento, todos os detalhes e altas frequências da imagem são reduzidos, assegurando que o *hash* não seja afetado pela diferença de dimensão ou pelos detalhes em relação ao *pixel* nas imagens duplicadas. Como a única informação relevante do *pixel* para o cálculo do *hash* é a intensidade luminosa, a imagem é convertida da escala RGB para a de cinza.

Feito isso, o *dHash* itera sobre cada par vizinho de *pixel* da imagem para calcular a diferença relativa entre o valor da intensidade de cor de cada *pixel*. A partir dessa iteração na matriz 9×8, é gerada uma matriz 8×8, que é chamada de **matriz de campo de diferença**, utilizada para calcular o valor de *hash* da imagem. As etapas necessárias para a obtenção dessa matriz são ilustradas na Figura 2.21.

Para cada par de *pixel* da matriz de campo de diferença, o valor do *hash* é gerado com base na seguinte condição: se o valor de intensidade do *pixel* da esquerda for maior que o do *pixel*

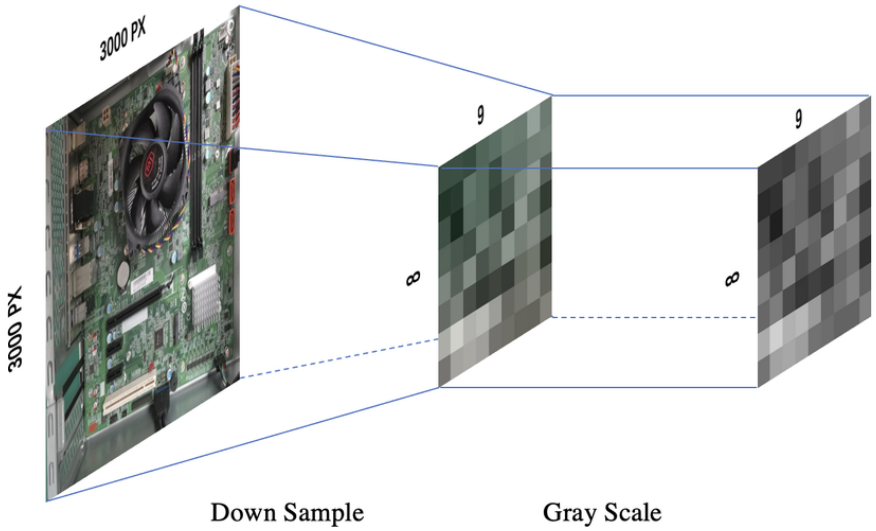

Figura 2.21: Etapas do dHash até a obtenção da matriz de campo de diferença (XIN ZIHAN ZHOU, 2023).

da direita, o valor para a posição referente a esse par de *pixels* é alterado para 01 na matriz resultante; caso contrário, é alterado para 0. A partir dessa matriz resultante, um *hash* de 64 bits é gerado para a imagem (MAHARANA, 2016).

2.10.5 Aumento Artificial de Dados

A técnica de aumento artificial de dados, conhecida como *data augmentation*, é amplamente utilizada para regularizar o treinamento de modelos de aprendizado de máquina. *Regularizar*, nesse contexto, significa aplicar métodos que ajudam a evitar o *overfitting*, uma situação em que o modelo se adapta excessivamente aos dados de treinamento, comprometendo sua capacidade de generalização para novos dados. O *data augmentation* cria novos exemplos de treinamento com base nos dados originais, ampliando a diversidade do conjunto de dados disponível (GÉRON, 2019).

Ampliar o conjunto de dados de treinamento permite que o modelo seja exposto a uma maior variedade de exemplos, contribuindo para melhorar sua capacidade de generalização e reduzir o risco de *overfitting*. Além disso, o uso do aumento artificial de dados também pode fortalecer a robustez do modelo diante de variações nos dados de entrada (GOODFELLOW; BENGIO; COURVILLE, 2016).

Essa abordagem é especialmente útil em problemas de classificação, como o abordado em nosso trabalho, pois as imagens envolvem alta dimensionalidade e apresentam uma ampla gama de variabilidades (GOODFELLOW; BENGIO; COURVILLE, 2016).

As alterações nos dados podem ser realizadas de diversas maneiras, incluindo mudanças no posicionamento, orientação, dimensões, iluminação, deslocamento e outros aspectos (GÉRON, 2019). Vale destacar que essa técnica é aplicada exclusivamente ao conjunto de treinamento, com o objetivo de regularizar o modelo. A Figura 2.22 apresenta um exemplo de aumento de dados. Alterações como luminosidade, cor, *zoom in*, *zoom out*, gradiente etc., foram aplicadas à imagem original para gerar novas imagens.

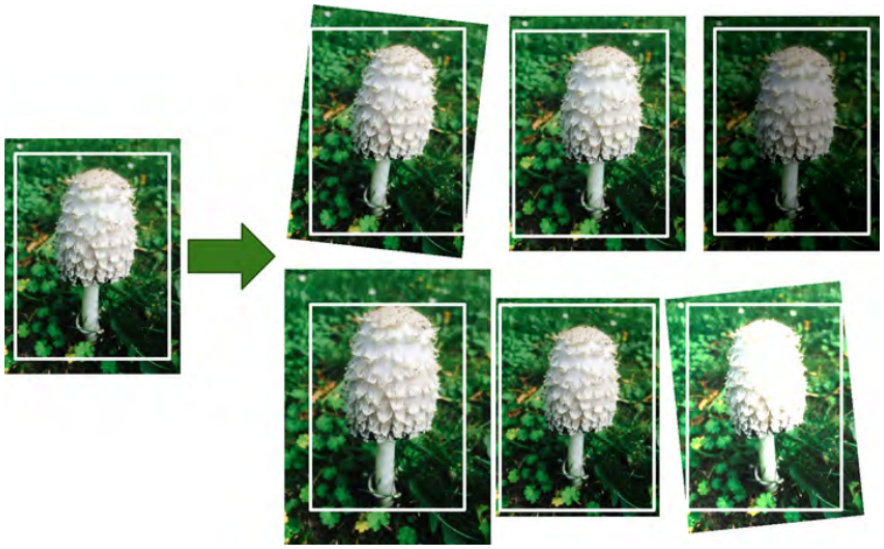

Figura 2.22: Exemplos de transformações aplicadas com o aumento de dados (GÉRON, 2019).

As alterações nos dados podem ser realizadas de diversas maneiras, incluindo mudanças no posicionamento, orientação, dimensões, iluminação, deslocamento e outros aspectos (GÉRON, 2019). Vale destacar que essa técnica é aplicada exclusivamente ao conjunto de treinamento, com o objetivo de regularizar o modelo.

2.10.6 Redução da Taxa de Aprendizagem em Razão do Platô

O platô em redes neurais artificiais refere-se a um estágio no processo de treinamento onde a diminuição da função de perda desacelera significativamente ou parece estagnar, mesmo com

iterações adicionais do algoritmo de otimização. Esse fenômeno pode ocorrer devido à presença de regiões quase-planas no espaço da função de perda, que são áreas onde os gradientes são muito pequenos (ARORA; LI; PANIGRAHI, 2022)

A técnica de redução da taxa de aprendizado em razão do platô é uma abordagem adaptativa para otimização de redes neurais que ajusta automaticamente a taxa de aprendizado quando o desempenho do modelo, medida pela função de perda ou pela métrica de validação, atinge um platô (SMITH, 2017). Essa estratégia é implementada ao monitorar o progresso do treinamento. Quando a melhoria cessa por um número fixo de épocas consecutivas, a taxa de aprendizado é reduzida, geralmente por um fator fixo, para permitir ajustes mais finos na busca pelo ótimo global (SMITH, 2017). Isso ajuda a superar dificuldades associadas a mínimos locais e regiões planas da função de perda.

2.10.7 *Early Stopping*

O*Early Stopping* é uma estratégia de regularização amplamente usada no treinamento de redes neurais para evitar o *overfitting*. Ela monitora o desempenho do modelo em um conjunto de validação durante o treinamento e interrompe o processo quando o desempenho começa a deteriorar-se ou estabilizar-se, indicando que o modelo está aprendendo padrões do ruído nos dados em vez de generalizar para novos exemplos (PRECHELT, 2002). Essa abordagem é especialmente útil para redes complexas, onde a identificação do momento ideal para encerrar o treinamento pode economizar recursos computacionais e otimizar a generalização (PRECHELT, 2002). O parâmetro de paciência (*patience*) é definido para determinar quantas épocas sem melhora no desempenho serão toleradas até que a parada seja acionada.

2.10.8 *Checkpoint*

A técnica de *Checkpoint* consiste em salvar o estado atual de um modelo durante o treinamento em intervalos regulares ou com base em critérios de desempenho, como, por exemplo, melhoria em métricas de validação (CHEN et al., 2014). Esse mecanismo visa prevenir perdas de progresso em caso de interrupções inesperadas e permitir a seleção do modelo com melhor de-

sempenho observado durante o treinamento. Os *checkpoints* são essenciais para trabalhos com modelos de aprendizado profundo de larga escala, pois reduzem significativamente os custos de computação em caso de falhas. Além disso, eles permitem retomar o treinamento a partir de um estado intermediário em vez de reiniciar do zero (EISENMAN et al., 2022).

2.10.9 Métricas de Avaliação

Para avaliar o desempenho de redes neurais, diferentes métricas são amplamente utilizadas. Elas permitem mensurar os modelos, oferecendo uma visão mais clara ao comparar seus resultados. É importante destacar que não há uma única maneira definitiva de determinar qual classificador apresenta o melhor desempenho (LUQUE et al., 2019). Em nosso trabalho, são utilizadas as métricas de acurácia, precisão, revocação e *f1-score*.

A matriz de confusão é uma ferramenta importante que ajuda a explorar os conceitos de cada métrica de avaliação do desempenho de redes neurais. A matriz de confusão é representada por uma matriz quadrada, utilizada para comparar as previsões do modelo com os valores reais (verdadeiros), como ilustrado na Figura 2.23. Os valores localizados na diagonal principal correspondem às previsões corretas, enquanto as demais posições representam os erros do modelo. Ela apresenta quatro categorias de rótulos:

- Verdadeiro Negativo (TN): quantidade de amostras corretamente classificadas como pertencentes à classe negativa.
- Verdadeiro Positivo (TP): quantidade de amostras corretamente classificadas como pertencentes à classe positiva.
- Falso Negativo (FN): número de amostras da classe positiva incorretamente classificadas como negativas.
- Falso Positivo (FP): número de amostras da classe negativa incorretamente classificadas como positivas.

Com isso, a partir dos parâmetros que compõem a matriz de confusão, é possível obter cada métrica de avaliação. Porém, é importante ressaltar que as métricas calculadas para tarefas

		Detectada	
		Sim	**Não**
Real	**Sim**	Verdadeiro Positivo (VP)	Falso Negativo (FN)
	Não	Falso Positivo (FP)	Verdadeiro Negativo (VN)

Figura 2.23: Modelo de Matriz de Confusão (RODRIGUES, 2019).

de multi-classificação são adaptadas para abranger todas as classes, diferenciando-se assim das métricas para tarefas de classificação binária. Para entender quais são essas adaptações para a multi-classificação e como são feitas, é necessário apresentar as métricas macro, micro e ponderadas.

As métricas macro avaliam o desempenho considerando cada classe individualmente e, em seguida, calculando a média das métricas de todas as classes. Essa abordagem atribui igual peso a todas as classes, independentemente do número de amostras de cada uma. Isso faz com que as métricas macro sejam ideais para problemas onde se deseja uma análise equilibrada entre classes, especialmente quando é importante avaliar o desempenho em classes menores, mesmo em conjuntos de dados desbalanceados. Porém, em cenários onde as classes são altamente desbalanceadas, o desempenho em classes majoritárias pode ser subestimado (LUQUE et al., 2019).

As métricas micro, por outro lado, combinam as contribuições de todas as classes ao somar os verdadeiros positivos, falsos positivos e falsos negativos antes de calcular as métricas de desempenho. Ao contrário das métricas macro, as métricas micro dão mais peso às classes maiores, já que suas contribuições dominam os valores agregados. Isso as torna adequadas para problemas onde o objetivo principal é maximizar o desempenho geral do modelo, priorizando as classes mais representadas. No entanto, elas podem mascarar o desempenho de classes menores, que têm menor influência no resultado final (GÉRON, 2019).

As métricas ponderadas são uma combinação entre as abordagens macro e micro. Elas calculam as métricas individuais para cada classe, mas ponderam essas métricas pelo número de amostras de cada classe antes de calcular a média final. Desse modo, as métricas ponderadas

conseguem equilibrar a análise de classes minoritárias e majoritárias, sendo úteis para conjuntos de dados desbalanceados onde é desejável refletir o impacto real das classes no desempenho geral (LUQUE et al., 2019). Em nosso trabalho, essas métricas foram utilizadas, devido à versatilidade e utilidade para conjuntos de dados desbalanceados, como é o caso do conjunto utilizado em nosso trabalho.

Nas fórmulas de cada métrica a seguir, a variável N representa o número total de classes:

- Acurácia: mede a proporção de previsões corretas em relação ao total de amostras avaliadas. É o resultado da razão total de acertos do modelo para a quantidade de todos os casos do teste, conforme a equação 2.13.

$$\text{Acurácia} = \frac{1}{N}\sum_{i=1}^{N}\frac{TP_i + TN_i}{TP_i + TN_i + FP_i + FN_i} \tag{2.13}$$

- Precisão: é a métrica utilizada para representar quantos casos de verdadeiros positivos são identificados corretamente pelo modelo, conforme a equação 2.14.

$$\text{Precisão} = \frac{1}{N}\sum_{i=1}^{N}\frac{TP_i}{TP_i + FP_i} \tag{2.14}$$

- Revocação (*recall*): também chamada de *sensibilidade*, é utilizada para representar as amostras positivas que foram corretamente rotuladas, conforme a equação 2.15.

$$\text{Revocação} = \frac{1}{N}\sum_{i=1}^{N}\frac{TP_i}{TP_i + FN_i} \tag{2.15}$$

- *F1-score*: é calculado a partir da média harmônica da precisão e sensibilidade, conforme a equação 2.16. O *f1-score* é especialmente útil em contextos onde há um desbalanceamento entre as classes, pois penaliza modelos que apresentam alta precisão, mas baixa revocação, ou vice-versa.

$$\textit{F1-score} = \frac{1}{N}\sum_{i=1}^{N}\frac{2 \times \text{Precisão} \times \text{Revocação}}{\text{Precisão} + \text{Revocação}} \tag{2.16}$$

2.11 Conclusão

Esse capítulo apresentou conceitos essenciais para fundamentar os objetivos propostos no trabalho. Foram explorados temas como, por exemplo, Inteligência Artificial, Aprendizado de Máquina, Redes Neurais Artificiais, Redes Neurais Convolucionais, *Transformers* e *Vision Transformers*, e definições de métodos e estratégias de implementação.

O Capítulo 3 apresenta e discute os trabalhos relacionados que fizeram uso de CNNs, ou *Transformers*, para a tarefa de classificação de imagens de doenças de pele.

Capítulo 3

Trabalhos Relacionados

Na literatura, trabalhos explorando o uso de CNNs e, mais recentemente, ViTs, na área da Saúde são comumente encontrados. As seguintes seções apresentam alguns trabalhos que fizeram uso de CNNs ou ViTs como solução para a tarefa de classificação de imagens. Por fim, na Seção 3.6 é realizada uma análise comparativa, apresentando uma visão crítica para cada trabalho relacionado.

3.1 *A Deep Learning Application for Psoriasis Detection*

O trabalho (MILANI et al., 2023) tem como foco a aplicação de Redes Neurais Convolucionais (CNNs) para detecção de psoríase por meio da tarefa de classificação binária de imagens. A pesquisa enfatiza a importância do diagnóstico rápido e preciso na melhoria da qualidade de vida dos pacientes, na redução de custos e no aumento da conscientização sobre a doença. O experimento visa comparar o desempenho de diferentes arquiteturas de CNN com o intuito de selecionar o modelo mais eficaz para detecção de lesões cutâneas associadas à psoríase.

Na metodologia, os autores coletaram um conjunto de dados composto por 2.260 imagens, incluindo 1.130 imagens de psoríase e 1.130 imagens de pele saudável. As imagens foram provenientes de diversas plataformas, incluindo *Google Images* para as imagens de pele saudável

e fontes dermatológicas certificadas disponibilizadas na Web, como o DermNet (DERMNETNZ, 2024), para as imagens de pele com psoríase. O estudo empregou técnicas de pré-processamento de dados, incluindo a validação cruzada K-*Fold* para particionamento de conjuntos de dados, para garantir o treinamento e a avaliação robustos dos modelos.

Os resultados indicaram que o modelo Inception-v3 superou as outras arquiteturas CNN, como ResNet50 (HE et al., 2016) e VGG19 (SIMONYAN, 2014), em todas as métricas de desempenho, conforme a Figura 3.1. O desempenho superior do Inception-v3 foi representado visualmente através de curvas ROC e matrizes de confusão, demonstrando sua eficácia na classificação precisa de imagens de psoríase. O estudo também destacou a importância do aumento de dados para melhorar a robustez do modelo e alcançar uma melhor generalização.

Modelos	**Acurácia média**	**Precisão média**	**Recall média**	**Especificidade média**	**F1-Score média**
Resnet50	73,4% ± 3,7	73,5% ± 4,4	73,8% ± 7,2	73% ± 7,0	73,4% ± 4,2
Resnet50 com aumento de dados	70,7% ± 2,4	68,2% ± 3,9	78,9% ± 4,9	62,5% ± 8,1	73% ± 1,7
VGG19	90,3% ± 1,2	89,5% ± 1,0	91,4% ± 1,5	89,2% ± 1,0	90,5% ± 1,2
VGG19 com aumento de dados	89,2% ± 0,3	87,9% ± 0,7	91,1% ± 0,8	87,4% ± 0,9	89,5% ± 0,3
Inception v3	97,5% ± 0,2	96,9% ± 0,4	98,2% ± 0,3	96,8% ± 0,4	97,5% ± 0,2
Inception v3 com aumento de dados	96,8% ± 0,1	96% ± 0,3	97,8% ± 0,4	95,9% ± 0,3	96,9% ± 0,1

Figura 3.1: Métricas gerais em termos de média e desvio padrão (MILANI et al., 2023).

A pesquisa ressalta o potencial das técnicas de aprendizagem profunda na área da dermatologia, principalmente para o diagnóstico da psoríase. As descobertas sugerem que a implementação de CNNs pode levar a processos de diagnóstico mais objetivos e eficientes, beneficiando, em última análise, o atendimento ao paciente. Em considerações finais, os autores recomendam a exploração de hiperparâmetros adicionais e o uso de novas técnicas de otimização para traba-

lhos futuros, a fim de aprimorar ainda mais o desempenho do modelo em aplicações do mundo real.

3.2 *Smart identification of psoriasis by images using convolutional neural networks: a case study in China*

O trabalho (ZHAO et al., 2020) se concentrou na detecção da psoríase usando um conjunto de dados estruturados derivados de imagens clínicas dermatológicas e registros médicos. Foi implementado um processo abrangente de filtragem e anotação de dados, onde as imagens foram verificadas por três dermatologistas experientes do Hospital Xiangya [1]. Esse processo garantiu que o conjunto de dados final fosse padronizado e confiável, composto por imagens que representavam com precisão diversas doenças de pele, incluindo psoríase e outras condições eritematosas semelhantes.

O conjunto de dados compreendeu nove categorias de doenças de pele, incluindo líquen plano, parapsoríase, lúpus eritematoso e eczema, que compartilham semelhanças visuais com a psoríase. As imagens foram cuidadosamente selecionadas para aumentar a credibilidade diagnóstica, pois foram verificadas por meio de exames anatomopatológicos e históricos médicos. Os autores enfatizaram a importância de ter um conjunto de dados diversificado para melhorar o desempenho de generalização do modelo preditivo.

Durante a fase de pré-processamento de dados, imagens não qualificadas foram removidas para criar um conjunto de dados limpo. Foram excluídas quatro categorias específicas de imagens: aquelas obscurecidas por tratamentos tópicos, partes especiais do corpo com lesões menores, lesões cobertas por pelos e imagens com exsudato excessivo. Esse rigoroso processo de filtragem teve como objetivo garantir que as imagens restantes fossem interpretáveis e adequadas para o treinamento do modelo preditivo.

Os resultados do estudo indicaram que o modelo de CNN Inception-v3 obteve o melhor

[1] Rua Xiangya, 410008, Changsha, Hunan, China.

desempenho na identificação de psoríase, AUC de $0,981 \pm 0,015$, superando outras arquiteturas como Xception (CHOLLET, 2017) e DenseNet121 (HUANG et al., 2017), conforme pode ser observado na Figura 3.2. Quando comparado com 25 dermatologistas, o sistema de IA demonstrou maior precisão $(0,96)$ e menor taxa de erros de diagnóstico $(0,19)$, demonstrando sua confiabilidade em ambientes clínicos. Além disso, o modelo manteve uma precisão global de $0,88$ em novos conjuntos de testes, confirmando as suas robustas capacidades de generalização e apoiando o seu potencial como uma ferramenta eficaz para melhorar a eficiência diagnóstica em dermatologia.

	AUC (95%CI)	Specificity	Sensitivity
InceptionV3 (One-stage)	0.966 ± 0.004	0.96	0.91
DenseNet121 (Two-stage)	0.954 ± 0.024	0.97	0.83
InceptionResNetV2 (Two-stage)	0.975 ± 0.022	0.97	0.95
Xception (Two-stage)	0.976 ± 0.022	0.98	0.93
InceptionV3 (Two-stage)	0.981 ± 0.015	0.98	0.92

Figura 3.2: AUC dos modelos para a detecção da psoríase (ZHAO et al., 2020).

Os autores concluem que o modelo baseado em rede neural convolucional desenvolvido para identificação da psoríase representa um avanço significativo no diagnóstico dermatológico, particularmente em áreas com acesso limitado a cuidados especializados. Os resultados promissores, incluindo alta precisão e capacidades robustas de generalização, sugerem que a IA pode efetivamente ajudar os dermatologistas a fazer diagnósticos oportunos e precisos. Além disso, a pesquisa destaca a importância de um conjunto de dados bem estruturado e de anotações especializadas no treinamento de modelos confiáveis, abrindo caminho para estudos futuros explorarem a integração da IA no diagnóstico de outras doenças de pele e na melhoria dos resultados gerais dos pacientes em dermatologia.

3.3 *Enhancing Skin Disease Classification Leveraging Transformer-based Deep Learning Architectures and Explainable AI*

O artigo (MOHAN et al., 2024) apresenta uma nova abordagem para a classificação de doenças de pele, aproveitando arquiteturas avançadas de aprendizagem profunda baseadas em *Transformers*, e concentrando-se especificamente no modelo DinoV2 (OQUAB et al., 2023) recentemente introduzido. O estudo visa melhorar a precisão e a eficiência do diagnóstico de doenças de pele, o que pode levar a tratamentos mais rápidos e eficazes por especialistas em pele. Ao utilizar um conjunto de dados aumentado de doenças de pele de 31 classes, os autores alcançaram uma notável precisão de teste de $96,48\%$ e um *f1-Score* de $0,9728$, marcando uma melhoria significativa em relação às métricas de classificação existentes.

A metodologia empregada nessa pesquisa envolveu uma avaliação abrangente de várias arquiteturas de *Transformers*, incluindo *Vision Transformers* e *Swin Transformers*, juntamente com redes neurais convolucionais tradicionais. Os autores conduziram experimentos em conjuntos de dados bem conhecidos, como Dermnet e HAM10000 (TSCHANDL; ROSENDAHL; KITTLER, 2018), para avaliar a robustez e generalização dos modelos propostos. Conforme as métricas apresentadas na Figura 3.3, os resultados demonstraram que o modelo DinoV2 superou os *benchmarks* anteriores, mostrando o seu potencial para uma classificação precisa de doenças de pele.

Authors	Year	Classes	Architecture	Accuracy	Precision	Recall	F1-Score
A. Rafay and W. Hussain [24]	2023	31	EfficentNetB2	0.8715	0.87	0.87	0.87
Proposed Work		31	ViT-Base	0.9235	0.9367	0.9370	0.9349
		31	Swin-Base	0.9326	0.9488	0.9516	0.9472
		31	**DinoV2-Base**	**0.9648**	**0.9755**	**0.9711**	**0.9728**

Figura 3.3: Métricas Resultantes (MOHAN et al., 2024).

Além do desempenho da classificação, o artigo enfatiza a importância de técnicas de *Explainable AI* (XAI), como GradCAM e SHAP, para melhorar a interpretabilidade das previsões do modelo. Essas ferramentas fornecem *insights* sobre os processos de tomada de decisão do sistema de IA, permitindo que os dermatologistas entendam os recursos que contribuem para os

diagnósticos. De acordo com os autores, a interpretabilidade é crucial para construir confiança nas aplicações de IA na área médica e pode ajudar os médicos a tomar decisões informadas em relação ao atendimento ao paciente.

Em conclusão, os autores reconhecem certas limitações no seu trabalho, como a complexidade computacional dos modelos e a generalização a diversos conjuntos de dados. Eles sugerem que pesquisas futuras devem se concentrar no desenvolvimento de arquiteturas leves para manter o desempenho da classificação e, ao mesmo tempo, reduzir os requisitos de recursos. No geral, o estudo destaca o potencial dos modelos baseados em *Transformers* para revolucionar o diagnóstico de doenças de pele, beneficiando, em última análise, tanto os profissionais de saúde como os pacientes, melhorando os resultados do tratamento e reduzindo os encargos financeiros para os sistemas de saúde.

3.4 *LesionAid: Vision Transformers-based Skin Lesion Generation and Classification*

O trabalho (KRISHNA et al., 2023) introduz o *framework LesionAid*, que utiliza *Vision Transformers* (ViT) e *Generative Adversarial Networks* (GAN) (GOODFELLOW et al., 2020) para a classificação de lesões cutâneas. O objetivo principal é enfrentar os desafios do desequilíbrio de classes em conjuntos de dados de lesões cutâneas, gerando imagens sintéticas, melhorando assim o conjunto de dados de treinamento. O *framework* é projetado para ser compatível com unidades computacionais de alto desempenho e de baixo custo.

A arquitetura do *framework* incorpora um codificador *Transformer* composto de blocos *Multi-Headed Self Attention* (MSA) e blocos *Multi-Layer Perceptron* (MLP), que trabalham juntos para processar sequências de incorporação de entrada. As formulações matemáticas fornecidas no artigo descrevem as operações realizadas nos componentes MSA e MLP, enfatizando o uso de conexões residuais e camadas de normalização.

De acordo com os autores, a preparação do conjunto de dados é crucial para o desempenho do *framework*. Por isso, utilizaram o conjunto de dados HAM10000, que contém 10.015 imagens

dermatoscópicas. As imagens de entrada são processadas em *patches*, que então são achatadas e projetadas em um espaço de dimensão inferior para criar *embeddings* de *patches*. Esse processo permite que o modelo aprenda e classifique com eficácia várias lesões cutâneas, alcançando precisões de treinamento e validação de $99,2\%$ e $97,4\%$, respectivamente.

Por fim, o artigo também discute a importância da *Explainable AI* (XAI) no contexto do *framework LesionAid*. A XAI visa tornar os modelos de IA mais interpretáveis, permitindo aos usuários compreender os processos de tomada de decisão por trás das previsões. Destaca-se também nesse estudo a implantação de uma aplicação web, que fornece uma interface amigável para análise de lesões cutâneas em tempo real, contribuindo, em última análise, para melhorar o atendimento dermatológico e os resultados dos pacientes.

3.5 *A Skin Disease Classification Model Based on DenseNet and ConvNeXt Fusion*

O trabalho (WEI et al., 2023) apresenta um modelo de classificação aprimorado que integra dois submodelos para melhorar o desempenho em tarefas de classificação de imagens de lesões cutâneas. O modelo é baseado na arquitetura ConvNeXt e foi ajustado para melhor extrair os recursos das imagens específicas do conjunto de dados coletado. Os autores enfatizam a importância de combinar os pontos fortes de diferentes submodelos para alcançar um sistema de classificação abrangente e preciso.

O modelo proposto utiliza uma série de blocos ConvNeXt aprimorados, que incluem várias camadas convolucionais, técnicas de normalização e operações de *pooling*. A arquitetura é projetada para agrupar de forma adaptativa os recursos extraídos e concatená-los para classificação. Os modelos foram pré-treinados no conjunto de dados ImageNet e ajustes foram feitos nos pesos do modelo para que se alinhassem com a nova estrutura.

As métricas de desempenho, como, por exemplo, acurácia, precisão, revocação e *f1-score* foram calculadas para avaliar a eficácia do modelo. Os resultados indicam que o modelo proposto supera vários modelos relevantes da literatura, incluindo DenseNet201 e ConvNeXt-L,

particularmente em cenários com conjuntos de dados desequilibrados ou tamanhos de amostra limitados, conforme os resultados apresentados pela Figura 3.4. As matrizes de confusão ilustram ainda mais as capacidades superiores de classificação do modelo em diferentes classes, conforme apresentado na Figura 3.5.

Model	Acne	Melasma	Rosacea	Nevus of Ota	Macro-Average
ResNet50	94.10	82.84	75.93	78.99	82.97
EfficientNet_B4	95.39	86.42	88.49	80.00	87.58
DenseNet201	95.62	84.81	90.00	85.25	88.92
ConvNeXt_L	95.94	87.80	89.66	86.67	90.02
Ours	98.12	93.98	94.91	93.10	95.03

Figura 3.4: *F1-Score*, em porcentagem, de cada modelo (WEI et al., 2023).

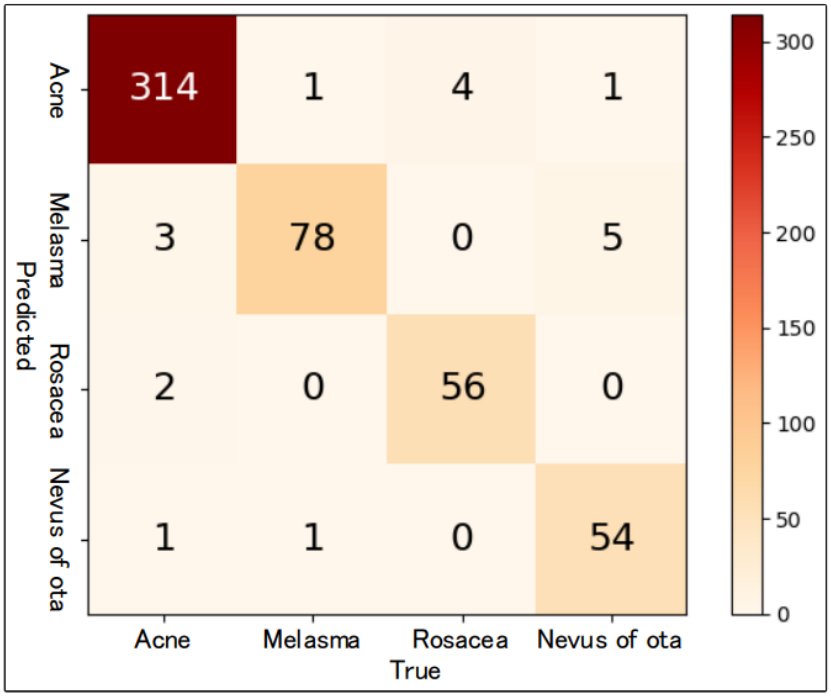

Figura 3.5: Matriz de confusão do modelo obtido (WEI et al., 2023).

Os autores concluem que o modelo proposto demonstra forte desempenho de classificação e capacidade de generalização, tornando-o adequado para diversas aplicações em reconhecimento de imagens. A pesquisa foi apoiada por financiamento do Programa Nacional de Pesquisa e Desenvolvimento da China e da Fundação de Medicina de Precisão Tsinghua [2].

[2]Distrito Haidian, 100084, Beijing, China.

3.6 Análise Comparativa

Assim como o trabalho (MILANI et al., 2023), nossa proposta explora a aplicação de técnicas de aprendizagem profunda para detecção de psoríase em imagens de pele. A metodologia e o escopo, entretanto, diferem. O artigo restringe seu foco à classificação binária de imagens, prevendo apenas se a imagem em questão está rotulada como saudável ou com psoríase. Esse fator diminui consideravelmente o nível de complexidade do problema, induzindo a obtenção de resultados otimistas. Além disso, o artigo apresenta a comparação entre três modelos de CNN: ResNet50 (HE et al., 2016), Inception-v3 (SZEGEDY et al., 2016b) e VGG19 (SIMONYAN, 2014), que são modelos clássicos, relativamente antigos quando em comparação com a data de publicação do artigo.

Outro estudo que se concentrou na tarefa de detecção da psoríase foi o (ZHAO et al., 2020), que seleciona um conjunto de dados de imagens dermatológicas abrangendo diversas doenças de pele com características semelhantes às da psoríase. Além disso, os autores também realizaram um refinamento durante o pré-processamento do conjunto de dados, removendo imagens que não eram qualificadas para a fase de treinamento. No entanto, apesar de os dados terem sido rotulados por especialistas da área, por não estarem disponíveis publicamente, a comunidade científica pode enfrentar limitações para reproduzir o estudo ou propor melhorias. Além disso, foi mencionado o uso de técnicas de aumento de dados, mas não ficou claro se essas técnicas foram restritas ao conjunto de treinamento. Caso sejam aplicadas aos conjuntos de validação ou teste, informações artificiais que não correspondem à distribuição real dos dados podem ser introduzidas. Isso comprometeria a avaliação da capacidade de generalização do modelo, que deve ser testada em dados inéditos. Assim, é essencial limitar o aumento de dados ao conjunto de treinamento.

O artigo (MOHAN et al., 2024) apresenta os resultados sobre o desempenho de modelos de ViT na tarefa de multiclassificação de imagens de doenças de pele. As imagens utilizadas para compor o conjunto de dados foram coletadas das bases de dados DermNet e HAM10000. Os autores não focam no diagnóstico da psoríase especificamente. Utilizam um conjunto de dados aumentado de 31 classes, propondo um desafio de multiclassificação para várias doenças de pele.

No entanto, os autores não clarificam se o conjunto de dados foi ampliado ou refinado. Apesar de os autores concluírem que os modelos ViT do experimento obtiveram um maior desempenho em relação à CNN EfficientNetB2, essa comparação entre CNNs e ViTs se diferencia da do presente trabalho, uma vez que este lida com um conjunto de dados específico de imagens de doenças de psoríase e doenças similares.

O trabalho (KRISHNA et al., 2023) desenvolve um *framework* chamado *LesionAid* com o objetivo de classificar lesões cutâneas no geral. Os autores fazem uso de *Vision Transformers* e enfrentam o desafio de lidar com o desbalanceamento no conjunto de dados. Eles exploram o desempenho de *Generative Adversarial Networks* (GANs), uma outra categoria de rede neural artificial. Além disso, os autores utilizam imagens de várias doenças de pele para compor o conjunto de dados. Entretanto, a arquitetura proposta não aborda suficientemente técnicas para superar o problema de *overfitting*, que é destacado como uma preocupação em múltiplas partes do texto.

Por fim, o trabalho (WEI et al., 2023) faz uso do modelo ConvNeXt pré-treinado a partir do conjunto de dados do ImageNet, com a finalidade de obter um modelo capaz de classificar de forma eficiente imagens de diferentes tipos de lesões cutâneas, como, por exemplo, acne, melasma, rosácea e nevo de Ota. Os autores construíram um novo modelo a partir da fusão dos modelos ConvNeXt e DenseNet, ao invés de apenas aplicar estratégias de transferência de aprendizado. Porém, os autores não chegam a realizar algum processo de refinamento nas imagens do conjunto de dados, como a remoção de imagens de pele coberta por cabelos ou desfocadas, o que pode dificultar a generalização do modelo.

3.7 Conclusão

Nesse capítulo foram apresentados os seguintes trabalhos que estão relacionados com essa monografia: (MILANI et al., 2023) (Seção 3.1), (ZHAO et al., 2020) (Seção 3.2), (MOHAN et al., 2024) (Seção 3.3), (KRISHNA et al., 2023) (Seção 3.4) e (WEI et al., 2023) (Seção 3.5). Além disso, foi realizada uma análise comparativa (Seção 3.6) onde foram discutidos os pontos de conexão com o presente trabalho, bem como as diferenças, críticas e pontos de melhoria dos

trabalhos relacionados.

No Capítulo 4 são apresentados os métodos propostos para a implementação do experimento que esse trabalho propõe.

Capítulo 4

Proposta de Métodos e Implementação

Para identificar o modelo com melhor desempenho na tarefa de detecção da psoríase, torna-se necessário definir, implementar e testar os métodos de multiclassificação de imagens que deverão ser utilizados. Nesse capítulo são apresentadas as bases de imagens utilizadas e o refinamento das imagens selecionadas (Seção 4.1); o particionamento do conjunto de dados (Seção 4.2) e o pré-processamento das imagens (Seção 4.3); os modelos e arquiteturas CNN e ViT selecionadas para a multiclassificação de imagens de psoríase e doenças similares (Seção 4.4); a transferência de aprendizado (Seção 4.4.1); e, por fim, as configurações do treinamento (Seção 4.5).

4.1 Obtenção das Imagens e Refinamento

A obtenção de imagens médicas apresentou-se como um desafio durante a etapa de construção da base de dados. Devido a questões éticas e à necessidade de proteger a privacidade dos pacientes retratados em imagens dermatológicas, é imprescindível obter autorizações e consentimentos explícitos. Essa exigência dificulta o acesso e o compartilhamento dessas imagens.

Algumas plataformas de dermatologia acessíveis na Web, como, por exemplo, a DermNet (DERMNETNZ, 2024), Dermis (DERMIS, 2022), DermaToWeb (WEB. . . , 2002), entre outras, disponibilizam as imagens já obtidas mediante consentimento prévio dos pacientes, para serem utilizadas para fins educativos e acadêmicos.

Nesse sentido, foi coletado o conjunto de imagens rotuladas como psoríase, dermatite, líquen

plano e pitiríase rosada, de diferentes áreas do corpo, a partir de repositórios de domínio público (DERMNETNZ, 2024; DERMIS, 2022; WEB..., 2002; DERMATOLOGY..., 2024; ATLAS..., 2017; FOR..., 2011). Todas as imagens coletadas possuem etiquetas de diagnóstico, validadas por dermatologistas certificados. A partir dessa coleta, foi possível obter um total de 3.248 imagens de pele doente.

As imagens coletadas que apresentavam lesões em áreas específicas como unhas e couro cabeludo foram desconsideradas. Isso foi feito porque essas imagens omitiam características visuais importantes da mancha da doença. Também foram removidas do conjunto as imagens sem foco ou borradas. A quantidade de imagens removidas totalizou 463, resultando em um conjunto com 2.781 imagens de pele com a doença devidamente etiquetadas.

Além disso, foram coletadas imagens de pele saudável, também de diversas áreas do corpo humano, para compor o conjunto de dados. As imagens de pele saudável foram coletadas de dados NTU (HUYNH et al., 2014a) do *Biometrics and Forensics Lab* (HUYNH et al., 2014b). O total de imagens coletadas dessa fonte resultou em um conjunto contendo 20 itens. Essa quantidade representa apenas uma pequena porção do conjunto de dados inteiro. O restante das imagens de pele saudável foi obtido manualmente a partir do Google Images (GOOGLE, 2024), resultando em um total de 1.176 imagens de pele saudável.

Um importante fator que deve ser considerado na construção de um conjunto de dados extraído de diferentes fontes públicas, é a possibilidade de existirem imagens duplicadas no conjunto. Para tratar isso, foi desenvolvido um *script*, utilizando Python, que implementa o algoritmo *dHash* (Seção 2.10.4) para detectar as imagens duplicadas em todo o conjunto de dados. Na execução do *script*, quando imagens duplicadas são detectadas, todas as duplicatas são mostradas na tela junto de seus respectivos nomes de arquivo. Dessa forma, foi possível validar a duplicidade e identificar qual das imagens deve permanecer no conjunto de dados. As duplicatas das imagens devem ser removidas do conjunto de dados.

A Tabela 4.1 apresenta, para cada rótulo (classe), a quantidade de imagens antes e depois dos refinamentos realizados. A Figura 4.1 apresenta o gráfico de distribuição de classes do conjunto de dados em sua totalidade.

Classe	Quantidade	Quantidade após refinamentos
Psoríase	1.438	1.176
Dermatite	1.411	1.137
Líquen plano	264	225
Pitiríase rosada	135	129
Saudável	1.254	1.176
Total	4.502	3.843

Tabela 4.1: Quantidade de imagens para cada classe, antes e após refinamentos.

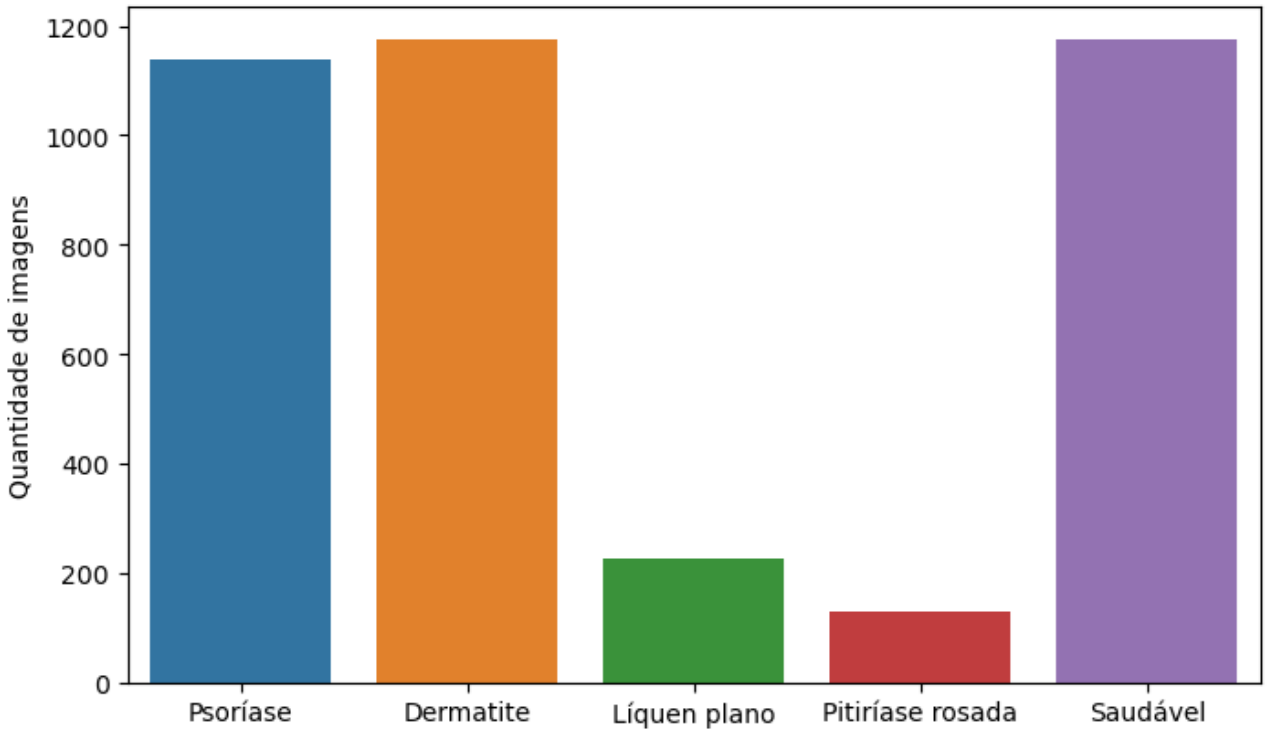

Figura 4.1: Distribuição do conjunto de dados por classe.

4.2 Particionamento do Conjunto de Dados

O particionamento do conjunto de dados seguiu a técnica de validação cruzada K-*Fold*, conforme descrito na Seção 2.10.3. A abordagem consistiu em embaralhar os dados e dividi-los em dois conjuntos estratificados, garantindo que as classes fossem representadas proporcionalmente em relação à distribuição original. Nessa divisão inicial, 80% do conjunto de dados destinou-se a participar da validação cruzada, representando os dados de treino e validação. A outra parte, 20%, foi separada exclusivamente para compor os dados de teste, que não participam da validação cruzada e nem da etapa de treino.

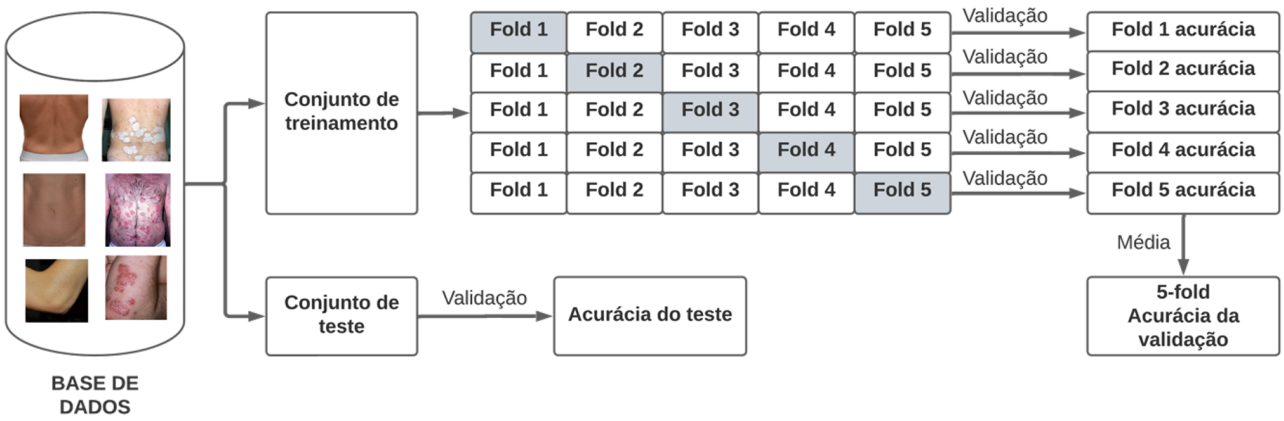

Figura 4.2: Validação cruzada K-*Fold* estratificada com k sendo cinco (MILANI et al., 2023).

Posteriormente, cinco grupos estratificados (*folds*) foram utilizados para treinamento, enquanto um foi reservado para validação, conforme a Figura 4.2. Durante cada iteração do processo de validação, um dos *folds* de treinamento é selecionado para validação, permitindo que o modelo seja avaliado em diferentes subconjuntos de dados. Esse método contribui para minimizar o risco de *overfitting* em apenas um conjunto específico e aumenta a confiabilidade e a precisão dos resultados obtidos (GÉRON, 2019).

4.3 Pré-Processamento das Imagens

Cada modelo selecionado possui uma especificação dos parâmetros de configuração para o pré-processamento das imagens. Dessa forma, para cada modelo, as imagens foram redimensionadas para um tamanho específico conforme a Tabela 4.2.

Modelo	**Redimensionamento**
Inception-v3	229×229
EfficientNetV2-L	480×480
ConvNeXt-L	232×232
ViT-L/16	242×242
MaxViT-T	224×224
DaViT-B	224×224

Tabela 4.2: Especificação do redimensionamento das imagens para cada modelo.

Além disso, todas as imagens tiveram suas intensidades de pixel normalizadas, dividindo-as pelo valor 255, de modo a colocá-las em uma escala entre 0 (zero) e 1 (um). A normalização

das imagens utilizando essa escala facilita o processamento pelas redes neurais, uma vez que os valores são ajustados para um intervalo menor, tornando os cálculos mais eficientes. A divisão por 255 é utilizada porque representa o valor máximo que um pixel pode assumir em imagens RGB ou em escala de cinza, já que cada canal de cor (vermelho, verde e azul) varia de 0 a 255 (CHOLLET, 2022).

O conjunto de dados destinado ao treinamento passou por um processo de aumento artificial de dados, conforme descrito na Seção 2.10.5. Essa técnica incluiu rotações incrementais das imagens em ângulos aleatórios entre 0° e 20°, além da aplicação de giros tanto horizontais quanto verticais.

4.4 Modelos e Treinamento

O processo de projetar uma rede neural profunda (DNN) envolve diversas decisões importantes, como a definição do número de camadas, a dimensão de cada camada, a quantidade de neurônios, o tipo de função de ativação e outros hiperparâmetros que podem impactar diretamente o desempenho do modelo. Embora algumas dessas escolhas sejam feitas com base em experimentos e testes empíricos, muitas dependem da experiência e do julgamento do projetista, fundamentados em conhecimentos teóricos e práticas consagradas na literatura científica. Por isso, o desenvolvimento de um novo modelo de DNN é uma tarefa complexa, que exige tempo e cuidado.

Nesse sentido, optou-se por utilizar a versão pré-treinada (a partir da base do ImageNet) de cada modelo, disponível nas bibliotecas TorchVision (MAINTAINERS; CONTRIBUTORS, 2016) e Timm (WIGHTMAN, 2019). A tabela 4.3 apresenta o tamanho e a quantidade de parâmetros de cada modelo, além de suas respectivas acurácias *top-1* e *top-5* na base de dados do ImageNet.

Os modelos foram selecionados com base em suas características específicas, buscando avaliar os resultados obtidos em diferentes configurações e levando em conta seu desempenho na competição ImageNet. Um fator comum entre eles é a ampla disponibilidade em bibliotecas de aprendizado profundo, como TorchVision e Timm, o que facilita sua implementação. Além

Modelo	Categoria	Tamanho (MB)	Parâmetros	Acurácia *top-1* (%)	Acurácia *top-5* (%)
Inception-v3	CNN	104	27,2M	77,3%	93,5%
EfficientNetV2-L	CNN	455	118,5M	85,8%	97,8%
ConvNeXt-L	CNN	755	197,8M	84,4%	97,0%
ViT-L/16	ViT	1.161	304,3M	79,7%	94,6%
MaxViT-T	ViT	118	30,9M	83,7%	96,7%
DaViT-B	ViT	352	88,0M	84,6%	96,9%

Tabela 4.3: Modelos selecionados e suas respectivas acurácias no conjunto de dados ImageNet (MAINTAINERS; CONTRIBUTORS, 2016; WIGHTMAN, 2019).

disso, a possibilidade de utilizar aprendizado por transferência se destaca, pois permite trabalhar com conjuntos de dados menores, além de reduzir significativamente o tempo e os custos associados ao treinamento. Esse aspecto será discutido com mais detalhes na próxima seção.

4.4.1 Transferência de Aprendizado

Treinar uma rede neural profunda completa para uma tarefa específica demanda elevados recursos computacionais e um grande volume de dados. Por isso, é comum recorrer a modelos pré-treinados, conforme discutido na Seção 2.10.1, que podem ser utilizados como ponto de partida ou como extratores de características para o problema em questão.

A estratégia definida para transferência de aprendizado (Seção 2.10.2) foi a de utilizar um modelo pré-treinado como extrator de características, devido às suas vantagens significativas de eficiência, simplicidade e redução do risco de *overfitting* (Seção 2.10.2).

4.5 Configurações do Treinamento

Durante o treinamento dos modelos, foi definido um critério para acompanhar as métricas de perda (*loss*) e acurácia (*accuracy*). Esse critério de otimização foi definido para o caso de o modelo apresentar melhorias na taxa de aprendizado (*learning rate*). Há várias técnicas destinadas a agilizar e otimizar a etapa de treinamento, especialmente quando o modelo não está apresentando um desempenho satisfatório. A seguir, serão apresentadas as configurações utilizadas em todos os treinamentos realizados.

- Otimizador: utilizou-se o otimizador *AdaMax* (PYTORCH, 2024a), que é uma versão do Adam mais robusta para grandes gradientes e valores de parâmetros elevados (Seção 2.5.3).

- *ReduceLROnPlateau* (PYTORCH, 2024b): essa função disponível na biblioteca PyTorch implementa a técnica de redução da taxa de aprendizado mediante platô, definida na Seção 2.10.6. A métrica utilizada para identificar o platô da taxa de aprendizagem é a perda (*loss*) do conjunto de validação. A redução aplicada tem um fator de 1e-3, e a paciência (*patience*) foi definida como 3, o que significa que o modelo espera por três épocas sem melhorias na métrica antes de diminuir a taxa de aprendizagem usando o fator.

- *Early Stopping*: vide definição na Seção 2.10.7. A acurácia do conjunto de validação é a métrica utilizada para avaliar o desempenho do modelo, e a paciência foi definida como 7 épocas.

- *Checkpoint* (PYTORCH, 2024c): vide definição na Seção 2.10.8. Utilizou-se as funções *save* (PYTORCH, 2024d) e *load_state_dict* (PYTORCH, 2024e) para salvar e carregar, respectivamente, os parâmetros do modelo.

- *Batch size*: vide definição na Seção 2.5.2. Utilizou-se um *batch* de tamanho 32 levando em consideração os recursos computacionais disponíveis e o tempo de treinamento.

- Épocas: vide definição na Seção 2.5.2. Para esse experimento, definiu-se 50 épocas.

4.6 Conclusão

Nesse capítulo foram apresentadas as bases de imagens utilizadas e o refinamento das imagens selecionadas (Seção 4.1); o particionamento do conjunto de dados (Seção 4.2); e o pré-processamento das imagens (Seção 4.3); os modelos e arquiteturas CNN e ViT selecionadas para a multi-classificação de imagens de psoríase e doenças similares (Seção 4.4); a transferência de aprendizado (Seção 4.4.1); e, por fim, as configurações do treinamento (Seção 4.5). No

Capítulo 5 são feitas algumas ponderações relativas às arquiteturas CNN e ViT escolhidas. São também apresentados os resultados obtidos, os principais achados decorrentes deles e é indicado o modelo com maior potencial para alcançar o objetivo proposto desse trabalho.

Capítulo 5

Resultados e Discussões

Nesse capítulo são apresentados os resultados obtidos por meio da aplicação das metodologias descritas no Capítulo 4. São comparadas as arquiteturas ViT e CNN, analisando-se as matrizes de confusão de cada modelo, bem como suas métricas de teste em termos de média ponderada e desvio padrão. Por fim, os resultados são discutidos, destacando os principais achados e indicando o modelo com maior potencial para cumprir com o objetivo desse trabalho.

5.1 Análise Comparativa dos Modelos

Observando as matrizes de confusão das Figuras 5.1 a 5.6, é possível notar a dificuldade que todos os modelos apresentaram para prever corretamente as imagens rotuladas com pitiríase rosada, que foi confundida principalmente com a dermatite. O modelo que mais cometeu esse erro foi o ViT-L/16, que chegou a classificar incorretamente 55% dos casos de pitiríase rosada como dermatite, enquanto que o modelo que menos cometeu esse equívoco foi o MaxViT-T com 10% de predições incorretas.

Outra dificuldade apresentada por todos os modelos foi a de classificar corretamente imagens de líquen plano, que também foi ocasionalmente confundido com dermatite. Novamente, o modelo ViT-L/16 foi o que mais se confundiu durante esse tipo de predição, errando 44% dos casos. Já o modelo que menos cometeu esse equívoco foi o EfficientNetV2-L, classificando incorretamente 9,3% dos casos.

A Tabela 5.1 apresenta as métricas (conforme as definições da Seção 2.10.9) de cada um dos seis modelos selecionados, ressaltando o fato de que, na tabela, os três primeiros (Inception-v3, EfficientNetV2-L e ConvNeXt-L) são modelos de CNN e os três restantes (ViT-L/16, MaxViT-T e DaViT-B), modelos de *Vision Transformer*. Além disso, a tabela apresenta para comparação o número de parâmetros de cada modelo, de modo a destacar a complexidade computacional exigida de cada um. É válido ressaltar ainda que, conforme citado na Seção 2.10.9, cada métrica foi calculada a partir da média ponderada considerando os resultados da validação cruzada K-*fold*.

Tabela 5.1: Métricas gerais em termos de média ponderada e desvio padrão.

Modelo	**Parâmetros**	**Acurácia**	**Precisão**	**Revocação**	**F1-*score***
Inception-v3	27,2M	$94,5\% \pm 0,5$	$94,6\% \pm 0,5$	$94,5\% \pm 0,5$	$94,5\% \pm 0,5$
EfficientNetV2-L	118,5M	$95,4\% \pm 0,3$	$95,5\% \pm 0,2$	$95,4\% \pm 0,3$	$95,4\% \pm 0,3$
ConvNeXt-L	197,8M	$95,6\% \pm 0,1$	$95,6\% \pm 0,2$	$95,6\% \pm 0,1$	$95,5\% \pm 0,1$
ViT-L/16	304,3M	$88,5\% \pm 6,2$	$88,0\% \pm 7,1$	$88,5\% \pm 6,2$	$87,8\% \pm 7,2$
MaxViT-T	30,9M	$96,0\% \pm 0,3$	$96,1\% \pm 0,3$	$96,0\% \pm 0,3$	$96,0\% \pm 0,3$
DaViT-B	88,0M	$96,4\% \pm 0,5$	$96,4\% \pm 0,5$	$96,4\% \pm 0,5$	$96,4\% \pm 0,5$

Dentre os modelos de CNN, ConvNeXt-L e EfficientNetV2-L alcançaram um *f1-score* de 95,4% e 95,5%, respectivamente. No entanto, esse desempenho foi obtido a custo de uma quantidade de parâmetros significativamente elevada de ambos os modelos (118,5 milhões e 197,8 milhões, respectivamente), o que pode indicar uma maior limitação quanto à sua implementação prática em ambientes com recursos limitados, já que a quantidade de parâmetros de uma rede neural pode influenciar diretamente no seu custo computacional, conforme citado na Seção 2.5. Por outro lado, o modelo Inception-v3 se apresenta como uma alternativa mais econômica em termos de recursos computacionais, dado que conta com apenas 27,2 milhões de parâmetros, ao custo de um desempenho de predição ligeiramente inferior, mas ainda notável, com um *f1-score* de 94,5%.

Dentre os modelos de ViT, MaxViT-T e DaViT-B alcançaram as métricas mais altas (acurácia, precisão, revocação e *f1-score*) ao custo de quantidades de parâmetros relativamente baixas (30,9 milhões e 88,0 milhões, respectivamente). O modelo ViT-L/16, apesar de sua grande con-

tagem de parâmetros (304,3 milhões), obteve o pior desempenho em comparação com os outros modelos, alcançando um *f1-score* inferior, de 87,8%. Esse desempenho pode ser justificado pelo fato de o modelo não ter sido bem ajustado para o conjunto de dados do ImageNet, conforme aponta (TAN; LE, 2021).

As métricas do ViT-L/16 apresentaram os maiores valores de desvio padrão. Isso pode ser justificado, conforme citado por (ALEXEY, 2020), pelo fato de os modelos ViT dependerem fortemente da etapa de pré-treinamento e serem sensíveis ao tamanho do conjunto de dados de teste. Isso atinge principalmente as variantes mais pesadas como o ViT-L/16, ao lidar com menores amostras de teste. Com exceção do ViT-L/16, as métricas de todos os modelos apresentaram valores de desvio padrão relativamente baixos, o que indica que as métricas obtidas são consistentes e possuem pouca variação.

5.2 Discussão e Conclusões

As dificuldades por parte de todos os modelos na predição para as imagens de líquen plano e pitiríase rosada, confundidas com a dermatite, podem ser explicadas pela semelhança das lesões avermelhadas da dermatite com as da pitiríase rosada e líquen plano. Entretanto, há também a possibilidade dos modelos não terem sido capazes de aprender para as amostras dessas classes, uma vez que são minoritárias no conjunto de dados. Uma análise mais robusta e detalhada sobre esse aspecto poderá ser abordada em trabalhos futuros.

De modo geral, comparando o desempenho dos CNNs e ViTs, é possível concluir que ambas as arquiteturas possuem potencial para tratar de problemas de multi-classificação de imagens. Com exceção do ViT-L/16, apresentam métricas superiores a 94%, um resultado promissor. Entretanto, é notável a diferença na quantidade de parâmetros dos ViT MaxViT-T e DaViT-B, que utilizam uma quantidade significativamente menor, quando comparados com os modelos de CNN EfficientNetV2-L e ConvNeXt-L.

Por isso, é possível concluir que, para a tarefa de classificação de imagens de pele com psoríase proposta nesse trabalho, os modelos de ViT recentes (MaxViT-T e DaViT-B) performam de forma ligeiramente superior quanto ao desempenho de predição, mas de forma significati-

vamente superior quando se trata de economia de parâmetros e de recursos computacionais. Esse fator pode ser explicado, conforme citado na Seção 2.8.4, pelo fato de ViTs utilizarem estruturas de atenção (*attention*) no lugar de camadas de convolução, característica principal das CNNs.

A atenção nos ViTs permite que eles capturem características contextuais e globais de maneira mais compacta e eficiente do que as CNNs, resultando em uma rede neural com menos parâmetros. Essa economia de parâmetros permite um melhor aproveitamento dos recursos computacionais durante o treinamento do modelo. Por isso, os ViTs potencialmente alcançam um desempenho equivalente ou até superior ao de modelos de CNN (KHAN et al., 2022).

Finalmente, é possível observar o desempenho superior dos modelos DaViT-B e MaxViT-T para a tarefa proposta, principalmente por ambos terem alcançado métricas de predição excepcionais, mesmo sendo modelos menores em comparação com o restante, com exceção do Inception-v3, mas que apresentou menores métricas de predição. Entretanto, entre esses dois modelos, o que menos cometeu erros na classificação da psoríase, especificamente, foi o DaViT-B. Apesar de ser ligeiramente mais pesado que o MaxViT-T, o DaViT-B acertou 97% das classificações de psoríase. Esse fator, somado a sua conquista de maior *f1-score* (96,4%), consolida o DaViT-B como o modelo com maior potencial para atuar como ferramenta de auxílio no diagnóstico da psoríase. Mesmo ao se deparar com lesões visualmente similares às da psoríase, DaViT-B se mostra capaz de alcançar excelentes resultados comparado aos outros modelos do experimento.

5.3 Conclusão

Nesse capítulo foram apresentados os resultados obtidos por meio da aplicação das metodologias descritas no Capítulo 4. Foram comparadas as arquiteturas ViT e CNN e os resultados foram discutidos, destacando os principais achados e indicando o modelo DaViT-B como o de maior potencial para atuar como ferramenta de auxílio no diagnóstico da psoríase.

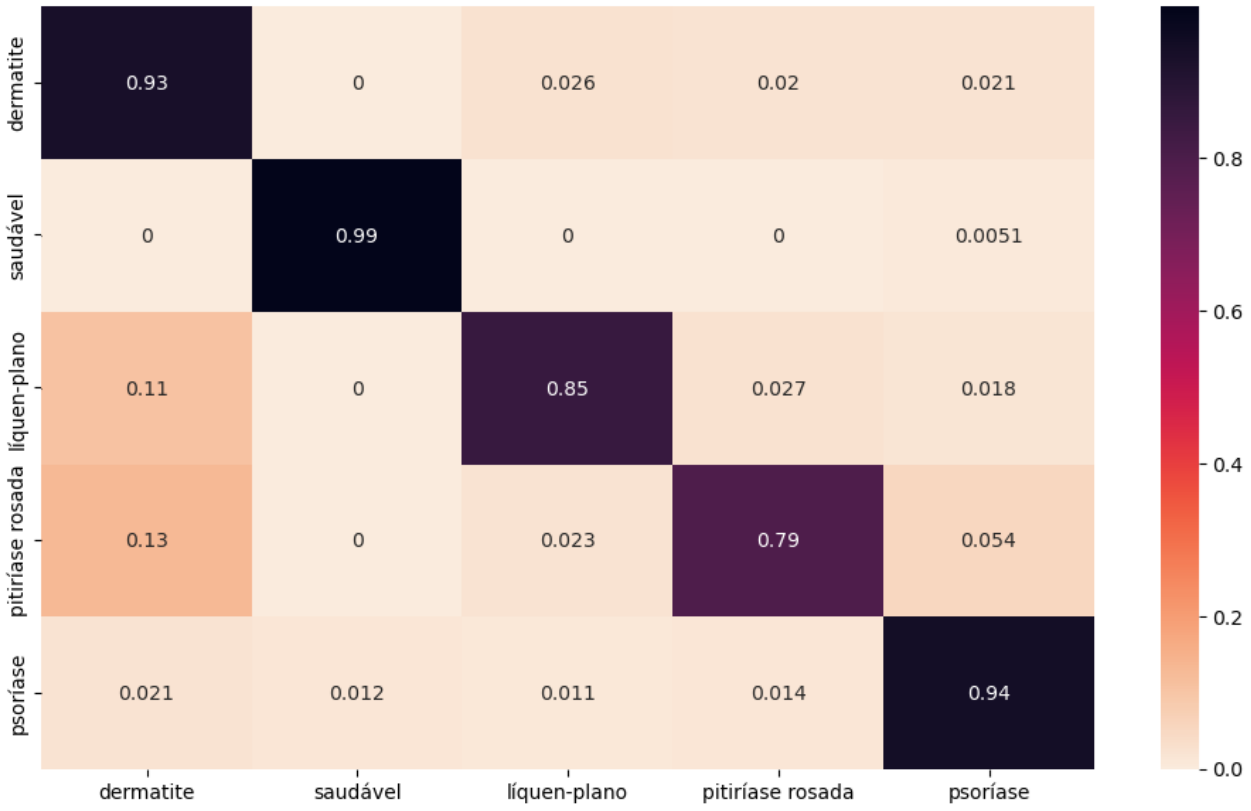

Figura 5.1: Matriz de confusão do modelo Inception-v3.

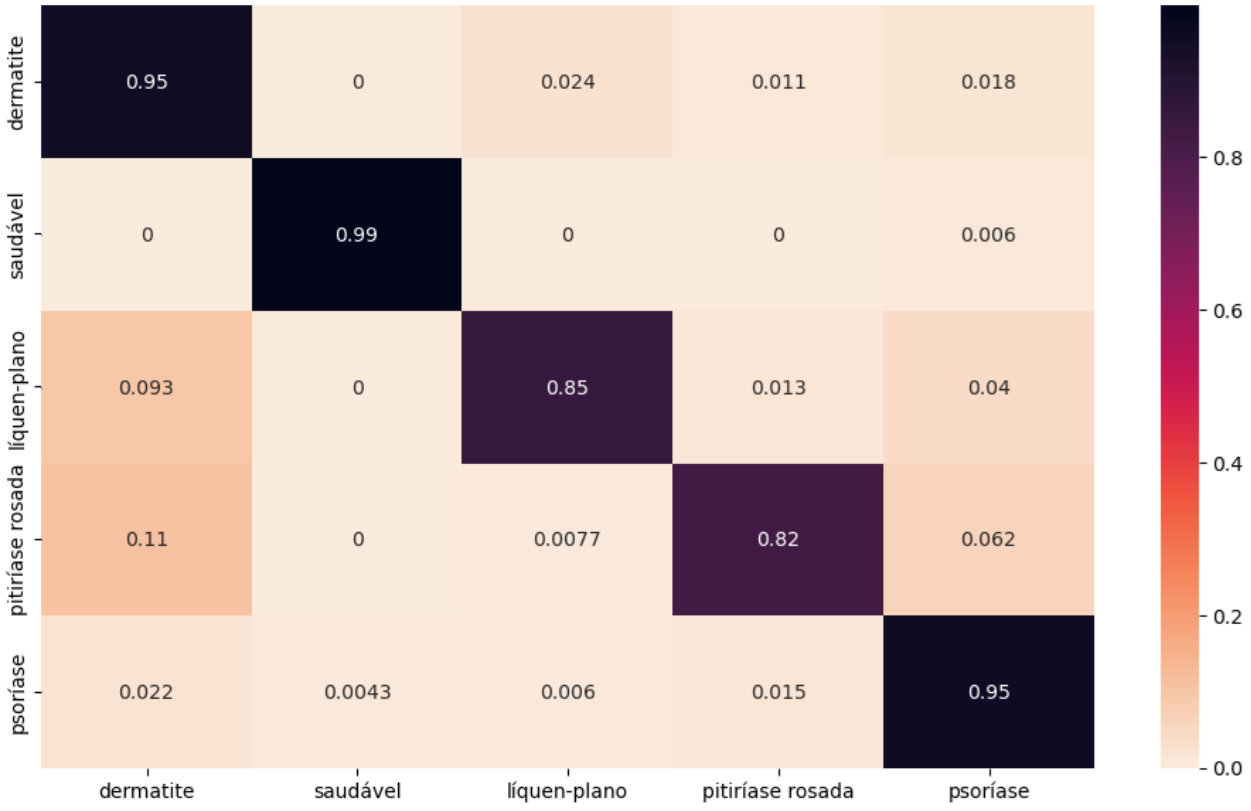

Figura 5.2: Matriz de confusão do modelo EfficientNetV2-L.

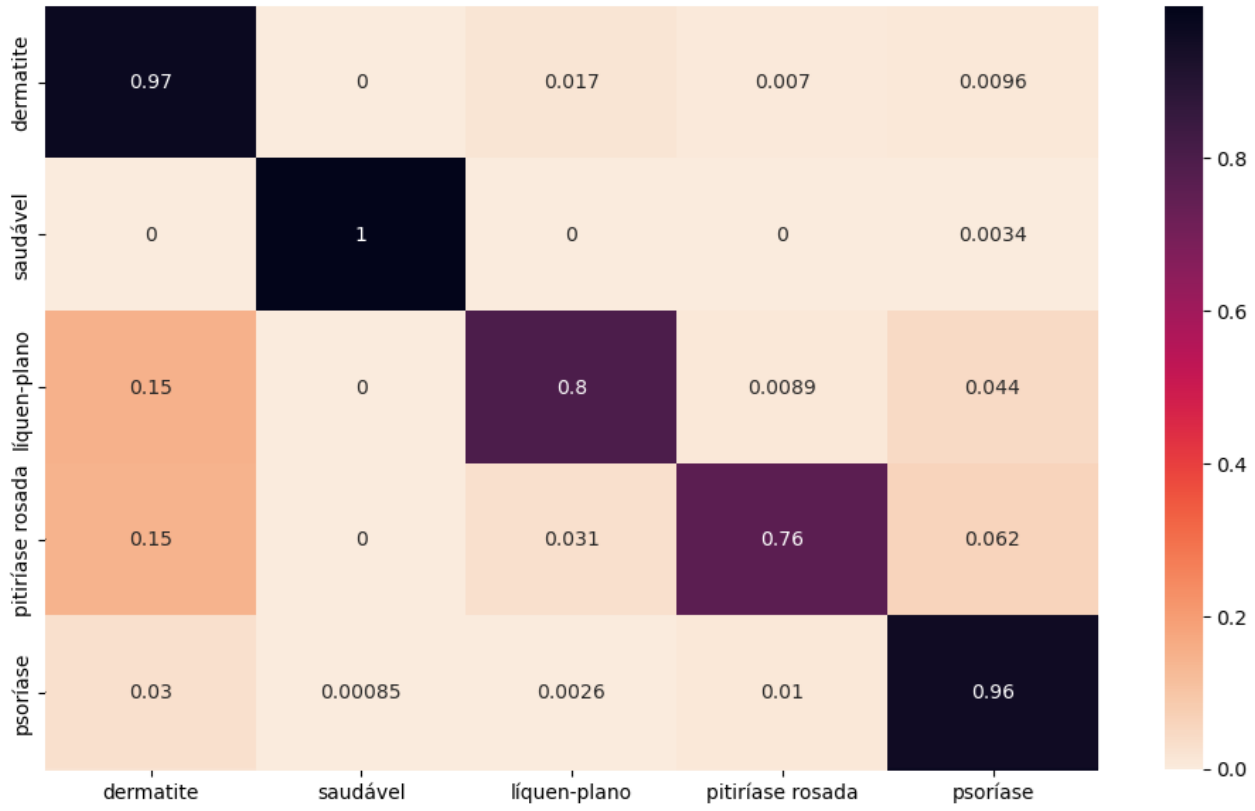

Figura 5.3: Matriz de confusão do modelo ConvNeXt-L.

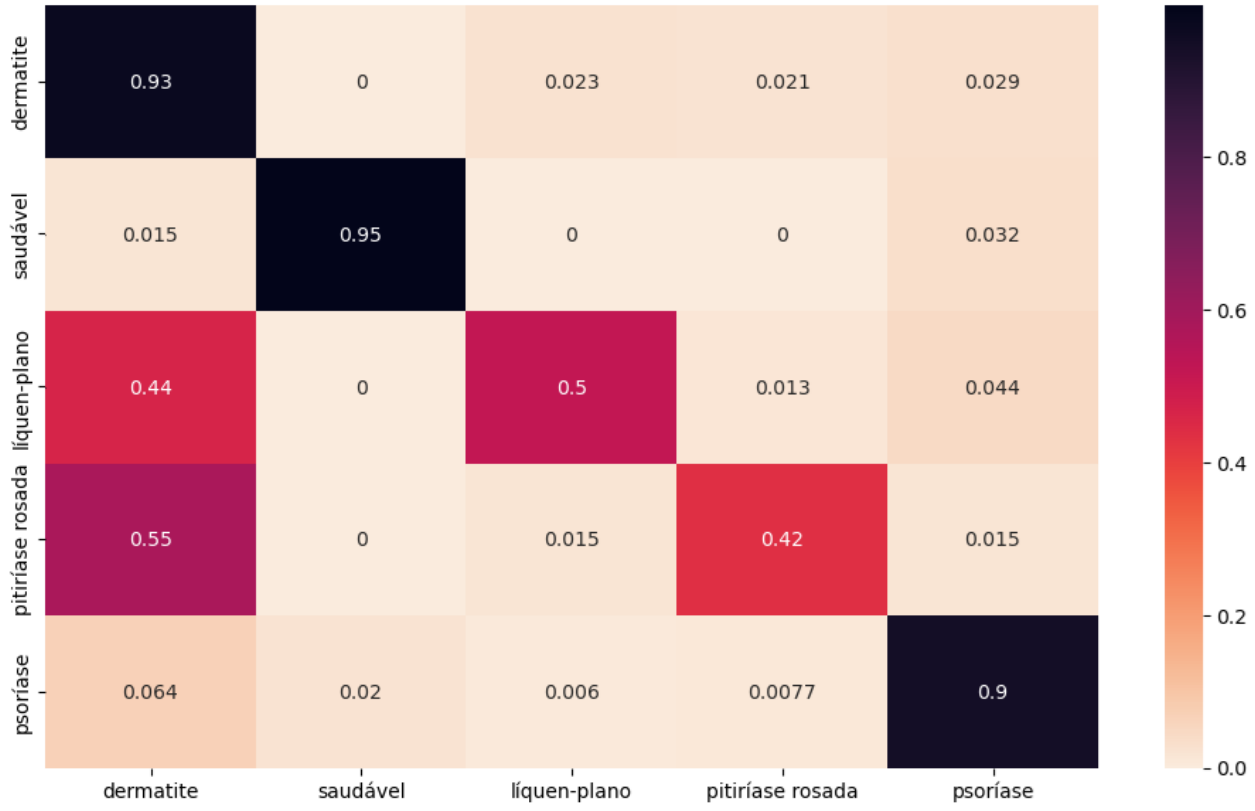

Figura 5.4: Matriz de confusão do modelo ViT-L/16.

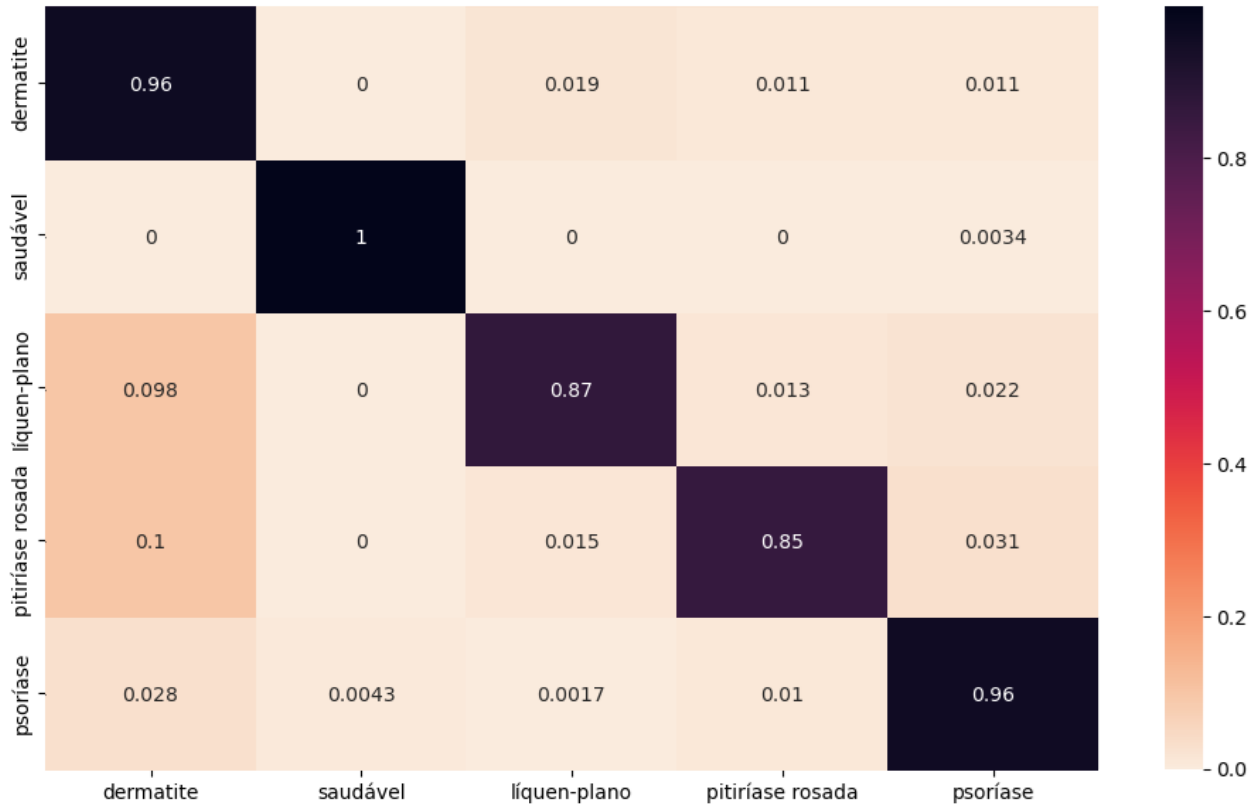

Figura 5.5: Matriz de confusão do modelo MaxViT-T.

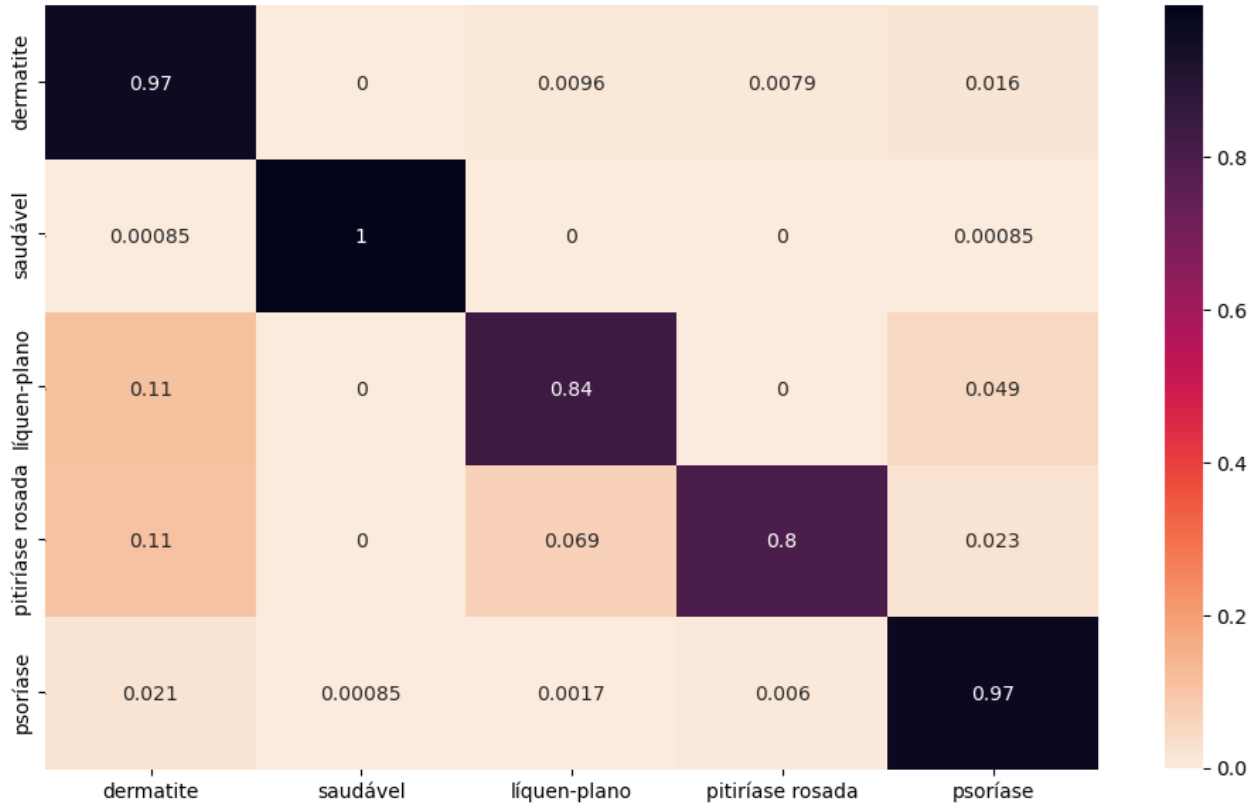

Figura 5.6: Matriz de Confusão do modelo DaViT-B.

Capítulo 6

Considerações Finais

Este trabalho teve como objetivo geral identificar o modelo de Aprendizado Profundo mais promissor para detecção da psoríase. Para isso, foram selecionadas três Redes Neurais Convolucionais (Inception-v3, EfficientNetV2-L e ConvNeXt-L) e três *Vision Transformers* (ViT-L/16, MaxViT-T e DaViT-B), todos pré-treinados a partir do conjunto de dados do ImageNet. Foram utilizadas imagens de pele com lesões causadas pelas seguintes doenças similares à psoríase: pitiríase rosada; dermatites; e líquen plano. O objetivo de cada modelo foi classificar as lesões de psoríase e aquelas similares a essa. Cada modelo foi submetido às etapas de treinamento e validação utilizando técnicas como a técnica de Validação Cruzada K-*Fold* (com $K = 5$), para otimizar a generalização.

O conjunto de dados de teste passou por uma etapa de validação. A partir dos resultados dessa validação, os modelos de CNN e ViT selecionados foram comparados. A comparação considerou a eficácia de cada modelo em classificar as lesões de psoríase e de doenças similares, com o propósito de também comparar a capacidade de discernimento das doenças visualmente similares à psoríase. Cada métrica de teste foi coletada a partir do cálculo de média ponderada, respeitando o resultado de cada iteração do K-*Fold*.

A conclusão da comparação é que ambas as categorias de modelos (CNN e ViT) apresentam um grande potencial para a resolução da tarefa proposta, com todos os modelos alcançando um f1-*score* superior a 94%, com exceção do ViT-L/16. No entanto, os modelos da categoria ViT MaxViT-T e DaViT-B se destacaram por alcançar resultados ligeiramente superiores, mesmo

sendo modelos significativamente mais leves do que as CNNs testadas. A partir disso, foi possível observar nas matrizes de confusão do MaxViT-T e DaViT-B que o modelo que menos cometeu erros na classificação da psoríase foi o DaViT-B. Isso constatou a sua eficiência na detecção da psoríase, sendo o modelo selecionado como o mais indicado para a tarefa.

A implementação de modelos como o DaViT-B em ambientes clínicos para auxiliar dermatologistas na detecção da psoríase é promissora, uma vez que o desempenho robusto do modelo em diferenciar psoríase de outras condições de pele visualmente similares demonstra seu potencial para reduzir diagnósticos equivocados. Isso favorece a detecção mais refinada da doença para que tratamentos mais precisos e eficazes sejam prescritos. Futuras investigações podem explorar a integração do DaViT-B com sistemas de imagem médica para a detecção de um leque maior de doenças de pele, ampliando sua aplicabilidade e garantindo maior relevância clínica.

6.1 Trabalhos Futuros

Os trabalhos futuros identificados durante a execução do trabalho são:

- Expandir o conjunto de dados de imagens, incluindo imagens de um leque maior de doenças de pele que se assemelhem à psoríase. Como exemplo, citamos o melanoma e outros tipos de câncer de pele;
- Realizar uma análise investigativa sobre os erros de predição apresentados por todos os modelos para as classes Líquen Plano e Pitiríase Rosada do conjunto de dados;
- Explorar mais a estratégia de transferência de aprendizado do ajuste fino do modelo pré-treinado, citado na Seção 4.4.1;
- Investigar a utilização do otimizador *Nesterov-accelerated Adaptative Moment Estimation* (Nadam) (DOZAT, 2016);
- Definir e coletar métricas de avaliação sobre o custo computacional de cada modelo, por exemplo a métrica de predições por segundo ou do tempo de uma predição;

- Verificar o desempenho, com ajuste fino, de modelos multimodais populares e modernos, como, por exemplo, Llama 3.2 (DUBEY et al., 2024), Gemini 1.5 (TEAM et al., 2024) e GPT-4o (HURST et al., 2024);
- Desenvolver um sistema que implemente o DaViT-B para o auxílio ao diagnóstico da psoríase e coletar métricas de teste desse sistema em ambientes de uso prático.

Referências Bibliográficas

ABRAHAM, A. *Artificial Neural Networks*. John Wiley & Sons, Ltd, 2005. ISBN 9780471497394. Disponível em: <https://onlinelibrary.wiley.com/doi/abs/10.1002/0471497398.mm421>.

ALEXEY, D. An image is worth 16x16 words: Transformers for image recognition at scale. *arXiv preprint arXiv: 2010.11929*, 2020.

ALHICHRI, H. et al. Classification of remote sensing images using efficientnet-b3 cnn model with attention. *IEEE access*, IEEE, v. 9, p. 14078–14094, 2021.

ALRFOU, K.; KORDIJAZI, A.; ZHAO, T. Computer vision methods for the microstructural analysis of materials: the state-of-the-art and future perspectives. *arXiv preprint arXiv:2208.04149*, 2022.

ARIFF, N. A. M.; ISMAIL, A. R. Study of adam and adamax optimizers on alexnet architecture for voice biometric authentication system. In: IEEE. *2023 17th International Conference on Ubiquitous Information Management and Communication (IMCOM)*. [S.l.], 2023. p. 1–4.

ARORA, S.; LI, Z.; PANIGRAHI, A. Understanding gradient descent on the edge of stability in deep learning. In: PMLR. *International Conference on Machine Learning*. [S.l.], 2022. p. 948–1024.

ASKARIZADEH, M. et al. Convex-concave programming: An effective alternative for optimizing shallow neural networks. *IEEE Transactions on Emerging Topics in Computational Intelligence*, IEEE, 2024.

ATLAS, G. P. *Explore the data*. 2024. <https://www.globalpsoriasisatlas.org/en/>. Acesso em: 12 de nov. de 2024.

ATLAS of Dermatology. 2017. <https://www.danderm.dk/atlas/>. Acesso em: 17/04/2024.

BALATO, N. et al. Effect of weather and environmental factors on the clinical course of psoriasis. *Occupational and Environmental Medicine*, BMJ Publishing Group Ltd, v. 70, n. 8, p. 600–600, 2013. ISSN 1351-0711. Disponível em: <https://oem.bmj.com/content/70/8/600>.

BASAVARAJ, K. H. et al. The role of drugs in the induction and/or exacerbation of psoriasis. *International Journal of Dermatology*, v. 49, n. 12, p. 1351–1361, 2010. Disponível em: <https://onlinelibrary.wiley.com/doi/abs/10.1111/j.1365-4632.2010.04570.x>.

BENGIO, Y.; DUCHARME, R.; VINCENT, P. A neural probabilistic language model. *Advances in neural information processing systems*, v. 13, 2000.

BRAGA, M. et al. Análise de sentimento com rede neural convolucional: uma investigação do fator motivacional da metodologia de aprendizagem criativa. In: *Anais do XLVIII Seminário Integrado de Software e Hardware*. Porto Alegre, RS, Brasil: SBC, 2021. p. 191–200. ISSN 2595-6205. Disponível em: <https://sol.sbc.org.br/index.php/semish/article/view/15822>.

BRANDRUP, F. et al. Psoriasis in monozygotic twins: variations in expression in individuals with identical genetic constitution. *Acta Derm. Venereol.*, v. 62, n. 3, p. 229–236, 1982.

BRASIL, A. *Estudo mostra que mais de 90% da população desconhecem a psoríase*. 2020. <https://agenciabrasil.ebc.com.br/saude/noticia/2020-11/estudo-mostra-que-mais-de-90-da-populacao-desconhecem-psoriase>. Acesso em: 29/04/2024.

BRINK, H.; RICHARDS, J. W.; FETHEROLF, M. *Real-world machine learning*. New York, NY: Manning Publications, 2016.

BURKE, R. Hybrid recommender systems: Survey and experiments. *User modeling and user-adapted interaction*, Springer, v. 12, p. 331–370, 2002.

CHEN, S. et al. Large-scale individual building extraction from open-source satellite imagery via super-resolution-based instance segmentation approach. *ISPRS Journal of Photogrammetry and Remote Sensing*, Elsevier, v. 195, p. 129–152, 2023.

CHEN, Z. et al. Numarck: machine learning algorithm for resiliency and checkpointing. In: IEEE. *SC'14: Proceedings of the International Conference for High Performance Computing, Networking, Storage and Analysis*. [S.l.], 2014. p. 733–744.

CHOLLET, F. Xception: Deep learning with depthwise separable convolutions. In: *Proceedings of the IEEE conference on computer vision and pattern recognition*. [S.l.: s.n.], 2017. p. 1251–1258.

CHOLLET, F. *Deep learning with python*. New York, NY: Manning Publications, 2022. 2-67 p.

CHOVATIYA, R.; SILVERBERG, J. I. Pathophysiology of atopic dermatitis and psoriasis: implications for management in children. *Children*, MDPI, v. 6, n. 10, p. 108, 2019.

CLEVERT, D.-A.; UNTERTHINER, T.; HOCHREITER, S. Fast and accurate deep network learning by exponential linear units (elus). arxiv 2015. *arXiv preprint arXiv:1511.07289*, 2020.

CODE, P. with. *Image Classification on ImageNet*. 2024. <https://paperswithcode.com/sota/image-classification-on-imagenet>. Acesso em: 28 de nov. de 2024.

COZMAN, F. G.; PLONSKI, G. A.; NERI, H. *Inteligência artificial: avanços e tendências*. Universidade de São Paulo. Instituto de Estudos Avançados, 2021. ISBN 9786587773131. Disponível em: <http://dx.doi.org/10.11606/9786587773131>.

DAS, S. Web Page, *Introdução à psoríase e às doenças descamativas*. [S.l.]: Manual MSD, 2023. <https://www.msdmanuals.com/pt/casa/dist%C3%BArbios-da-pele/psor%C3%ADase-e-dist%C3%BArbios-descamativos/introdu%C3%A7%C3%A3o-%C3%A0-psor%C3%ADase-e-%C3%A0s-doen%C3%A7as-descamativas>.

DASH, M. et al. A cascaded deep convolution neural network based cadx system for psoriasis lesion segmentation and severity assessment. *Applied Soft Computing*, 2020.

DENG, J. et al. Imagenet: A large-scale hierarchical image database. In: IEEE. *2009 IEEE conference on computer vision and pattern recognition*. [S.l.], 2009. p. 248–255.

DERMATOLOGIA, S. B. de. *Psoríase pode se associar a outras doenças*. 2016. <https://www.sbd.org.br/psoriase-pode-se-associar-a-outras-doencas/>. Acesso em: 29/04/2024.

DERMATOLOGIA, S. B. de. *O que é a psoríase?* 2023. <https://www.sbd.org.br/psoriase/>. Acesso em: 29/04/2024.

DERMATOLOGY Atlas. 2024. <https://atlasdermatologico.com.br/>. Acesso em: 19/04/2024.

DERMATOLOGY, L. *What is psoriasis?* 2024. <https://luminederm.com/condition/psoriasis/>. Acesso em: 21 de nov. de 2024.

DERMIS. *Dermatology Information System*. 2022. <https://www.dermis.net/>. Acesso em: 26/04/2024.

DERMNETNZ. *The worlds leading free dermatology website*. 2024. <https://dermnetnz.org/>. Acesso em: 29/04/2024.

DESAI, C. Comparative analysis of optimizers in deep neural networks. *International Journal of Innovative Science and Research Technology*, v. 5, n. 10, p. 959–962, 2020.

DHAR, S. et al. Psoriasis in children: An insight. *Indian Journal of Dermatology*, Medknow, v. 56, n. 3, p. 262, 2011. ISSN 0019-5154. Disponível em: <http://dx.doi.org/10.4103/0019-5154.82477>.

DING, M. et al. Davit: Dual attention vision transformers. In: SPRINGER. *European conference on computer vision*. [S.l.], 2022. p. 74–92.

DOGO, E. M. et al. A comparative analysis of gradient descent-based optimization algorithms on convolutional neural networks. In: IEEE. *2018 international conference on computational techniques, electronics and mechanical systems (CTEMS)*. [S.l.], 2018. p. 92–99.

DOZAT, T. Incorporating nesterov momentum into adam. 2016.

DUBEY, A. et al. The llama 3 herd of models. *arXiv preprint arXiv:2407.21783*, 2024.

EISENMAN, A. et al. {Check-N-Run}: A checkpointing system for training deep learning recommendation models. In: *19th USENIX Symposium on Networked Systems Design and Implementation (NSDI 22)*. [S.l.: s.n.], 2022. p. 929–943.

ELSHAMY, R. et al. Improving the efficiency of rmsprop optimizer by utilizing nestrove in deep learning. *Scientific Reports*, Nature Publishing Group UK London, v. 13, n. 1, p. 8814, 2023.

EYRE, R. W.; KRUEGER, G. G. Response to injury of skin involved and uninvolved with psoriasis, and its relation to disease activity: Koebner and 'reverse' koebner reactions. *British Journal of Dermatology*, Oxford University Press (OUP), v. 106, n. 2, p. 105–116, fev. 1982. ISSN 1365-2133. Disponível em: <http://dx.doi.org/10.1111/j.1365-2133.1982.tb00924.x>.

FOGAÇA, E. C. R. et al. Sistema de detecção de crianças em situações de perigo em piscinas usando deep learning e iot. Universidade do Estado do Amazonas, 2022.

FOR health professionals and public. 2011. <http://www.hellenicdermatlas.com/en/>. Acesso em: 16/04/2024.

FOUNDATION, P. S. *Python*. 2024. <https://www.python.org/>. Acesso em: 06 de dez. de 2024.

GELFAND, J. M. et al. The risk of mortality in patients with psoriasis: Results from a population-based study. *Archives of Dermatology*, American Medical Association (AMA), v. 143, n. 12, dez. 2007. ISSN 0003-987X. Disponível em: <http://dx.doi.org/10.1001/archderm.143.12.1493>.

GERING, G. B.; CIARELLI, P. M.; SALLES, E. O. Identificação de sentimento em voz por meio da combinação de classificações intermediárias dos sinais em excitação, valência e quadrante. In: SBC. *Anais do XIX Simpósio Brasileiro de Computação Aplicada à Saúde*. [S.l.], 2019. p. 152–163.

GÉRON, A. *Mãos À Obra: Aprendizado de Máquina com Scikit-Learn E TensorFlow*. [S.l.: s.n.], 2019.

GONÇALVES, A. R. Máquina de vetores suporte. v. 21, 2015.

GOODFELLOW, I.; BENGIO, Y.; COURVILLE, A. *Deep Learning*. [S.l.]: MIT Press, 2016. <http://www.deeplearningbook.org>.

GOODFELLOW, I. et al. Generative adversarial networks. *Communications of the ACM*, ACM New York, NY, USA, v. 63, n. 11, p. 139–144, 2020.

GOOGLE. 2024. Acesso em: 17 de maio de 2024. Disponível em: <https://www.google.com/imghp?hl=pt-BR>.

GRUS, J. *Data science do zero*. [S.l.]: Alta books Rio d Janeiro, 2016. v. 1.

GUIMARãES, A. J. et al. Using fuzzy neural networks to improve prediction of expert systems for detection of breast cancer. *Anais do XV Encontro Nacional de Inteligência Artificial e Computacional*, p. 799–810, 2018.

GURNEY, K. *An Introduction to Neural Networks*. Taylor & Francis, 1997. (An Introduction to Neural Networks). ISBN 9781857285031. Disponível em: <https://books.google.com.br/books?id=HOsvllRMMP8C>.

HAMMOND, K. *Pratical artificial intelligence for dummies*. [S.l.]: Narrative Science, 2015. 1–30 p.

HARRIS, C. R. et al. Array programming with NumPy. *Nature*, Springer Science and Business Media LLC, v. 585, n. 7825, p. 357–362, set. 2020. Disponível em: <https://doi.org/10.1038/s41586-020-2649-2>.

HAYKIN, S. *Neural networks: a comprehensive foundation*. [S.l.]: Prentice Hall PTR, 1998.

HE, K. et al. Deep residual learning for image recognition. In: *Proceedings of the IEEE conference on computer vision and pattern recognition*. [S.l.: s.n.], 2016. p. 770–778.

HENDRYCKS, D.; GIMPEL, K. Gaussian error linear units (gelus). *arXiv preprint arXiv:1606.08415*, 2016.

HENSELER, E. C. T. Psoriasis of early and late onset: Characterization of two types of psoriasis vulgaris. *Journal of the American Academy of Dermatology*, Elsevier BV, v. 13, n. 3, p. 450–456, set. 1985. ISSN 0190-9622. Disponível em: <http://dx.doi.org/10.1016/s0190-9622(85)70188-0>.

HO, J. et al. Axial attention in multidimensional transformers. *arXiv preprint arXiv:1912.12180*, 2019.

HUANG, G. et al. Densely connected convolutional networks. In: *Proceedings of the IEEE conference on computer vision and pattern recognition*. [S.l.: s.n.], 2017. p. 4700–4708.

HUNTER, J. D. Matplotlib: A 2d graphics environment. *Computing in Science & Engineering*, IEEE COMPUTER SOC, v. 9, n. 3, p. 90–95, 2007.

HURST, A. et al. Gpt-4o system card. *arXiv preprint arXiv:2410.21276*, 2024.

HURWITZ, J.; KIRSCH, D. *Machine Learning For Dummies*. [S.l.]: John Wiley & Sons, 2018. 75 p.

HUYNH, N. Q. et al. A preliminary report on a full-body imaging system for effectively collecting and processing biometric traits of prisoners. In: *2014 IEEE Symposium on Computational Intelligence in Biometrics and Identity Management (CIBIM)*. [S.l.: s.n.], 2014.

HUYNH, N. Q. et al. A preliminary report on a full-body imaging system for effectively collecting and processing biometric traits of prisoners. In: IEEE. *2014 IEEE Symposium on Computational Intelligence in Biometrics and Identity Management (CIBIM)*. [S.l.], 2014. p. 167–174.

JANKOVI, S. et al. Risk factors for psoriasis: A case–control study. *The Journal of Dermatology*, v. 36, n. 6, p. 328–334, 2009. Disponível em: <https://onlinelibrary.wiley.com/doi/abs/10.1111/j.1346-8138.2009.00648.x>.

JI YIYI WANG, D. Z. B. Automatic detection and evaluation of nail psoriasis based on deep learning: a preliminary application and exploration. *SPIE International Conference on Computer Application and Information Security*, 2022.

KENTON, J. D. M.-W. C.; TOUTANOVA, L. K. Bert: Pre-training of deep bidirectional transformers for language understanding. In: MINNEAPOLIS, MINNESOTA. *Proceedings of naacL-HLT*. [S.l.], 2019. v. 1, p. 2.

KHAN, S. et al. Transformers in vision: A survey. *ACM computing surveys (CSUR)*, ACM New York, NY, v. 54, n. 10s, p. 1–41, 2022.

KHAN, S. et al. *A Guide to Convolutional Neural Networks for Computer Vision*. Springer International Publishing, 2018. ISSN 2153-1064. ISBN 9783031018213. Disponível em: <http://dx.doi.org/10.1007/978-3-031-01821-3>.

KINGMA, D.; BA, J. Adam: A method for stochastic optimization. *International Conference on Learning Representations*, 12 2014.

KOWALCZYK, W.; SZLÁVIK, Z.; SCHUT, M. C. The impact of recommender systems on item-, user-, and rating-diversity. In: SPRINGER. *International Workshop on Agents and Data Mining Interaction*. [S.l.], 2011. p. 261–287.

KRISHNA, G. S. et al. Lesionaid: vision transformers-based skin lesion generation and classification. *arXiv preprint arXiv:2302.01104*, 2023.

KRIZHEVSKY, A.; SUTSKEVER, I.; HINTON, G. E. Imagenet classification with deep convolutional neural networks. *Communications of the ACM*, AcM New York, NY, USA, v. 60, n. 6, p. 84–90, 2017.

LAZAR, A. P.; ROENIGK, H. H. Acquired immunodeficiency syndrome (aids) can exacerbate psoriasis. *Journal of the American Academy of Dermatology*, Elsevier BV, v. 18, n. 1, p. 144, jan. 1988. ISSN 0190-9622. Disponível em: <http://dx.doi.org/10.1016/s0190-9622(88)80053-7>.

LECUN, Y.; DENKER, J.; SOLLA, S. Optimal brain damage. *Advances in neural information processing systems*, v. 2, 1989.

LI, Y. Research and application of deep learning in image recognition. In: IEEE. *2022 IEEE 2nd international conference on power, electronics and computer applications (ICPECA)*. [S.l.], 2022. p. 994–999.

LIN, G.-S. et al. Instance segmentation based on deep convolutional neural networks and transfer learning for unconstrained psoriasis skin images. *Applied Sciences*, MDPI, v. 11, n. 7, p. 3155, 2021.

LIU, Z. et al. Swin transformer: Hierarchical vision transformer using shifted windows. In: *Proceedings of the IEEE/CVF international conference on computer vision*. [S.l.: s.n.], 2021. p. 10012–10022.

LIU, Z. et al. A convnet for the 2020s. In: *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*. [S.l.: s.n.], 2022. p. 11976–11986.

LUQUE, A. et al. The impact of class imbalance in classification performance metrics based on the binary confusion matrix. *Pattern Recognition*, Elsevier, v. 91, p. 216–231, 2019.

MAAS, A. L. et al. Rectifier nonlinearities improve neural network acoustic models. In: ATLANTA, GA. *Proc. icml*. [S.l.], 2013. v. 30, n. 1, p. 3.

MADURANGA, P.; NANDASENA, D. Mobile-based skin disease diagnosis system using convolutional neural networks (cnn). *International Journal of Image, Graphics and Signal Processing*, v. 14, p. 47–57, 06 2022.

MAHARANA, A. *Application of Digital Fingerprinting: Duplicate Image Detection*. Dissertação (Mestrado) — National Institute of Technology Rourkela, 2016.

MAINTAINERS, T.; CONTRIBUTORS. *TorchVision: PyTorch's Computer Vision library*. [S.l.]: GitHub, 2016. <https://github.com/pytorch/vision>.

MANAV. *Introduction to Convolutional Neural Networks (CNN)*. 2024. <https://www.analyticsvidhya.com/blog/2021/05/convolutional-neural-networks-cnn/>. Acesso em: 22 de nov. de 2024.

MANJUSHA, V. S. L. Detect /remove duplicate images from a dataset for deep learning. *Journal of Positive School Psychology*, v. 6, p. 606–609, 2022.

MCKINNEY Wes. Data Structures for Statistical Computing in Python. In: WALT Stéfan van der; MILLMAN Jarrod (Ed.). *Proceedings of the 9th Python in Science Conference*. [S.l.: s.n.], 2010. p. 56 – 61.

MEHMOOD, F.; AHMAD, S.; WHANGBO, T. K. An efficient optimization technique for training deep neural networks. *Mathematics*, MDPI, v. 11, n. 6, p. 1360, 2023.

MIKOLOV, T. Efficient estimation of word representations in vector space. *arXiv preprint arXiv:1301.3781*, v. 3781, 2013.

MILANI, A. et al. A deep learning application for psoriasis detection. *Anais do Encontro Nacional de Inteligência Artificial e Computacional (ENIAC)*, p. 315–329, 2023.

MOHAN, J. et al. Enhancing skin disease classification leveraging transformer-based deep learning architectures and explainable ai. *arXiv preprint arXiv:2407.14757*, 2024.

MOON, C.-I. et al. Optimization of psoriasis assessment system based on patch images. *Scientific Reports*, Nature Publishing Group, v. 11, n. 1, p. 1–13, 2021.

MOURA ALESSANDRA MARINS COELHO, M. d. F. O. B. Luiza Rosa de. Detecção automática de anomalias oculares utilizando redes neurais convolucionais. *Anais do VIII Encontro Nacional de Computação dos Institutos Federais*, 2021.

NAIR, V.; HINTON, G. E. Rectified linear units improve restricted boltzmann machines. In: *Proceedings of the 27th international conference on machine learning (ICML-10)*. [S.l.: s.n.], 2010. p. 807–814.

NASSIRI, K.; AKHLOUFI, M. Transformer models used for text-based question answering systems. *Applied Intelligence*, Springer, v. 53, n. 9, p. 10602–10635, 2023.

OLESCKI, G. Detecção de tromboembolia pulmonar utilizando redes neurais convolucionais e extração de características. *Anais do XXI Simpósio Brasileiro de Computação Aplicada à Saúde*, p. 381–391, 2021.

ONG, K. L. et al. Maxmvit-mlp: Multiaxis and multiscale vision transformers fusion network for speech emotion recognition. *IEEE Access*, IEEE, 2024.

OQUAB, M. et al. Dinov2: Learning robust visual features without supervision. *arXiv preprint arXiv:2304.07193*, 2023.

ORR, G. B.; MÜLLER, K.-R. *Neural networks: tricks of the trade*. [S.l.]: Springer, 1998.

PAN, S. J.; YANG, Q. A survey on transfer learning. *IEEE Transactions on Knowledge and Data Engineering*, v. 22, n. 10, p. 1345–1359, 2010.

PAPP, K. et al. Canadian guidelines for the management of plaque psoriasis: Overview. *Journal of Cutaneous Medicine and Surgery*, SAGE Publications, v. 15, n. 4, p. 210–219, jul. 2011. ISSN 1615-7109. Disponível em: <http://dx.doi.org/10.2310/7750.2011.10066>.

PASZKE, A. et al. Pytorch: An imperative style, high-performance deep learning library. In: *Advances in Neural Information Processing Systems 32*. Curran Associates, Inc., 2019. p. 8024–8035. Disponível em: <http://papers.neurips.cc/paper/9015-pytorch-an-imperative-style-high-performance-deep-learning-library.pdf>.

PEDREGOSA, F. et al. Scikit-learn: Machine learning in Python. *Journal of Machine Learning Research*, v. 12, p. 2825–2830, 2011.

PRECHELT, L. Early stopping-but when? In: *Neural Networks: Tricks of the trade*. [S.l.]: Springer, 2002. p. 55–69.

PYTORCH. 2024. Acesso em: 29 de outubro de 2024. Disponível em: <https://pytorch.org/docs/stable/generated/torch.optim.Adamax.html>.

PYTORCH. 2024. Acesso em: 28 de outubro de 2024. Disponível em: <https://pytorch.org/docs/stable/generated/torch.optim.lr_scheduler.ReduceLROnPlateau.html>.

PYTORCH. 2024. Acesso em: 30 de outubro de 2024. Disponível em: <https://pytorch.org/docs/stable/checkpoint.html>.

PYTORCH. 2024. Acesso em: 04 de novembro de 2024. Disponível em: <https://pytorch.org/docs/stable/generated/torch.save.html\#torch.save>.

PYTORCH. 2024. Acesso em: 04 de novembro de 2024. Disponível em: <https://pytorch.org/docs/stable/generated/torch.nn.Module.html\#torch.nn.Module.load_state_dict>.

QUINTERO-ROJAS, J.; GONZáLEZ, J. D. Use of convolutional neural networks in smartphones for the identification of oral diseases using a small dataset. *Revista Facultad De Ingeniería*, Nature Publishing Group, 2021.

RADFORD, A. Improving language understanding by generative pre-training. 2018.

RAHMAN, N. *What is the benefit of using average pooling rather than max pooling?* 2018. <https://www.quora.com/What-is-the-benefit-of-using-average-pooling-rather-than-max-pooling/answer/Nouroz-Rahman>. Acesso em: 23 de nov. de 2024.

REDDI, S. J.; KALE, S.; KUMAR, S. On the convergence of adam and beyond. *arXiv preprint arXiv:1904.09237*, 2019.

ROBBINS, H.; MONRO, S. A stochastic approximation method. *The annals of mathematical statistics*, JSTOR, p. 400–407, 1951.

RODRIGUES, A. P.; TEIXEIRA, R. Desvendando a psoríase. *RBAC*, v. 41, n. 4, p. 303–309, 2009.

RODRIGUES, A. P.; TEIXEIRA, R. M. Desvendando a psoríase. *RBAC*, 2009.

RODRIGUES, V. *Métricas de Avaliação: acurácia, precisão, recall... quais as diferenças?* 2019. <https://vitorborbarodrigues.medium.com/m%C3%A9tricas-de-avalia%C3%A7%C3%A3o-acur%C3%A1cia-precis%C3%A3o-recall-quais-as-diferen%C3%A7as-c8f05e0a513c>. Acesso em: 02 de dez. de 2024.

ROMITI, R. et al. Psoriasis in childhood and adolescence. *Anais brasileiros de dermatologia*, SciELO Brasil, v. 84, p. 9–20, 2009.

ROSENBLATT, F. The perceptron: a probabilistic model for information storage and organization in the brain. *Psychological review*, American Psychological Association, v. 65, n. 6, p. 386, 1958.

ROSLAN, R. et al. Evaluation of psoriasis skin disease classification using convolutional neural network. *IAES International Journal of Artificial Intelligence (IJ-AI)*, v. 9, p. 349, 06 2020.

RUDER, S. An overview of gradient descent optimization algorithms. *ArXiv*, 2016. Disponível em: <https://api.semanticscholar.org/CorpusID:17485266>.

RUMELHART, D. E.; HINTON, G. E.; WILLIAMS, R. J. Learning representations by back-propagating errors. *nature*, Nature Publishing Group UK London, v. 323, n. 6088, p. 533–536, 1986.

RUSSAKOVSKY, O. et al. ImageNet Large Scale Visual Recognition Challenge. *International Journal of Computer Vision (IJCV)*, v. 115, n. 3, p. 211–252, 2015.

RUSSELL, S.; NORVIG, P. *Artificial intelligence*. 3. ed. Upper Saddle River, NJ: Pearson, 2009.

SEVILLE, R. Psoriasis and stress. *British Journal of Dermatology*, v. 97, n. 3, p. 297–302, 1977. Disponível em: <https://onlinelibrary.wiley.com/doi/abs/10.1111/j.1365-2133.1977.tb15186.x>.

SHAH, Y. et al. Deep learning model-based multimedia forgery detection. In: IEEE. *2020 Fourth International Conference on I-SMAC (IoT in Social, Mobile, Analytics and Cloud)(I-SMAC)*. [S.l.], 2020. p. 564–572.

SHANTHI, K.; MANIMEKALAI, S. Cervical cancer detection using ensemble neural network algorithm with stochastic gradient descent (sgd) optimizer. *SN Computer Science*, Springer, v. 5, n. 8, p. 1–12, 2024.

SILVA, B. M. da et al. Quando parece mas não é: Pitiríase rosea um diagnóstico diferencial de tinea corporis.

SILVA, G. et al. Cardiac arrhythmia detection in ecg signals using graph convolutional network. *Anais do XXII Simpósio Brasileiro de Computação Aplicada à Saúde*, p. 25–35, 2022.

SILVA, S. R. e. *Modelagem de oxigênio dissolvido utilizando redes neurais artificiais*. Dissertação (Mestrado) — Instituto Federal de Goiás, 2014.

SILVA, T. Um estudo comparativo entre algoritmos de aprendizagem de máquina supervisionados para predição de solução de reclamações no procon. Centro Universitário Christus-Unichristus, 2021.

SILVERBERG, J. I. Public health burden and epidemiology of atopic dermatitis. *Dermatologic clinics*, Elsevier, v. 35, n. 3, p. 283–289, 2017.

SIMONYAN, K. Very deep convolutional networks for large-scale image recognition. *arXiv preprint arXiv:1409.1556*, 2014.

SMITH, L. N. Cyclical learning rates for training neural networks. In: IEEE. *2017 IEEE winter conference on applications of computer vision (WACV)*. [S.l.], 2017. p. 464–472.

SOBREIRO, V. A. et al. Uma estimação do valor da commodity de açúcar utilizando redes neurais artificiais. *Pesquisa & Desenvolvimento em Engenharia de Produção*, v. 7, p. 36–52, 2008.

SOUZA, E. et al. Análise histológica comparativa entre epitélios saudáveis e doentes com psoríase e líquen plano. *Rev. Ciênc. Saúde Nova Esperança.[Internet]*, p. 1–15, 2016.

SOUZE, G. G. B. *Deep Learning para a Detecção e Classificação de Pneumonia por Radiografias do Tórax*. Dissertação (Mestrado) — Universidade Federal do Maranhão, 2018.

STEFANINI, M. et al. From show to tell: A survey on deep learning-based image captioning. *IEEE transactions on pattern analysis and machine intelligence*, IEEE, v. 45, n. 1, p. 539–559, 2022.

STEWART JULIA BATTAGLINI-SABETTA, L. M. A. F. Hypocalcemia-induced pustular psoriasis of von zumbusch. *Annals of Internal Medicine*, v. 100, n. 5, p. 677–680, 1984. PMID: 6546844. Disponível em: <https://www.acpjournals.org/doi/abs/10.7326/0003-4819-100-5-677>.

SZEGEDY, C. et al. *Inception-v4, Inception-ResNet and the Impact of Residual Connections on Learning*. 2016.

SZEGEDY, C. et al. Rethinking the inception architecture for computer vision. In: *Proceedings of the IEEE conference on computer vision and pattern recognition*. [S.l.: s.n.], 2016. p. 2818–2826.

TAN, M.; LE, Q. Efficientnet: Rethinking model scaling for convolutional neural networks. In: PMLR. *International conference on machine learning*. [S.l.], 2019. p. 6105–6114.

TAN, M.; LE, Q. Efficientnetv2: Smaller models and faster training. In: PMLR. *International conference on machine learning*. [S.l.], 2021. p. 10096–10106.

TEAM, G. et al. Gemini 1.5: Unlocking multimodal understanding across millions of tokens of context. *arXiv preprint arXiv:2403.05530*, 2024.

TSCHANDL, P.; ROSENDAHL, C.; KITTLER, H. The ham10000 dataset, a large collection of multi-source dermatoscopic images of common pigmented skin lesions. *Scientific data*, Nature Publishing Group, v. 5, n. 1, p. 1–9, 2018.

TU, Z. et al. Maxvit: Multi-axis vision transformer. In: SPRINGER. *European conference on computer vision*. [S.l.], 2022. p. 459–479.

VASWANI, A. Attention is all you need. *Advances in Neural Information Processing Systems*, 2017.

WASKOM, M. L. seaborn: statistical data visualization. *Journal of Open Source Software*, The Open Journal, v. 6, n. 60, p. 3021, 2021. Disponível em: <https://doi.org/10.21105/joss.03021>.

WASSERMAN, P. D.; SCHWARTZ, T. Neural networks. ii. what are they and why is everybody so interested in them now? *IEEE expert*, IEEE, v. 3, n. 1, p. 10–15, 1988.

WEB docente de Dermatologia. 2002. <http://dermatoweb.udl.es/>. Acesso em: 22/04/2024.

WEI, M. et al. A skin disease classification model based on densenet and convnext fusion. *Electronics*, MDPI, v. 12, n. 2, p. 438, 2023.

WIGHTMAN, R. *PyTorch Image Models*. [S.l.]: GitHub, 2019. <https://github.com/rwightman/pytorch-image-models>.

WISSTEIN, E. W. *Convolution*. 2024. <https://mathworld.wolfram.com/Convolution.html>. MathWorld–A Wolfram Web Resource. Acesso em: 28/04/2024.

WOOD, T. *What is the Softmax Function?* 2018. <https://deepai.org/machine-learning-glossary-and-terms/softmax-layer>. Acesso em: 23 de nov. de 2024.

WU, Z. et al. Studies on different cnn algorithms for face skin disease classification based on clinical images. *IEEE Access*, v. 7, p. 66505–66511, 2019.

XIN ZIHAN ZHOU, Y. X. Y. Scene separation & data selection: Temporal segmentation algorithm for real-time video stream analysis. *CEUR Workshop Proceedings*, 07 2023.

YEUNG, H. et al. Psoriasis severity and the prevalence of major medical comorbidity: A population-based study. *JAMA Dermatology*, American Medical Association (AMA), v. 149, n. 10, p. 1173, out. 2013. ISSN 2168-6068. Disponível em: <http://dx.doi.org/10.1001/jamadermatol.2013.5015>.

ZHAO, S. et al. Smart identification of psoriasis by images using convolutional neural networks: a case study in china. *Journal of the European Academy of Dermatology and Venereology*, Wiley Online Library, v. 34, n. 3, p. 518–524, 2020.

ZHOU, R.; KHEMMARAT, S.; GAO, L. The impact of youtube recommendation system on video views. In: . [S.l.: s.n.], 2010. p. 404–410.