UNIVERSIDADE FEDERAL DE UBERLÂNDIA
FACULDADE DE ENGENHARIA ELÉTRICA
PROGRAMA DE PÓS-GRADUAÇÃO EM ENGENHARIA ELÉTRICA

BIANCA BERTOLDO DE OLIVEIRA

DETECÇÃO DE MATURIDADE ÓSSEA UTILIZANDO ALGORITMO YOLO E REDES NEURAIS CONVOLUCIONAIS

Uberlândia

2025

BIANCA BERTOLDO DE OLIVEIRA

DETECÇÃO DE MATURIDADE ÓSSEA UTILIZANDO ALGORITMO YOLO E REDES NEURAIS CONVOLUCIONAIS

Dissertação apresentada ao Programa de Pós-Graduação em Engenharia Elétrica da Universidade Federal de Uberlândia, como parte dos requisitos para a obtenção do título de Mestre em Engenharia Elétrica.

Área de concentração: Processamento da Informação

Orientador: Prof. Dr. Antônio Cláudio Paschoarelli Veiga

Coorientadora: Profa. Dra. Milena Bueno Pereira Carneiro

Uberlândia

2025

Ficha Catalográfica Online do Sistema de Bibliotecas da UFU
com dados informados pelo(a) próprio(a) autor(a).

O48 2025

Oliveira, Bianca Bertoldo de, 1998-
Detecção de Maturidade Óssea utilizando algoritmo YOLO e Redes Neurais Convolucionais [recurso eletrônico] / Bianca Bertoldo de Oliveira. - 2025.

Orientador: Antônio Cláudio Paschoarelli Veiga.
Coorientadora: Milena Bueno Pereira Carneiro.
Dissertação (Mestrado) - Universidade Federal de Uberlândia, Pós-graduação em Engenharia Elétrica.
Modo de acesso: Internet.
DOI http://doi.org/10.14393/ufu.di.2025.661
Inclui bibliografia.
Inclui ilustrações.

1. Engenharia elétrica. I. Veiga, Antônio Cláudio Paschoarelli, 1963-, (Orient.). II. Carneiro, Milena Bueno Pereira,1980-, (Coorient.). III. Universidade Federal de Uberlândia. Pós-graduação em Engenharia Elétrica. IV. Título.

CDU: 621.3

Bibliotecários responsáveis pela estrutura de acordo com o AACR2:
Gizele Cristine Nunes do Couto - CRB6/2091
Nelson Marcos Ferreira - CRB6/3074

BIANCA BERTOLDO DE OLIVEIRA

DETECÇÃO DE MATURIDADE ÓSSEA UTILIZANDO ALGORITMO YOLO E REDES NEURAIS CONVOLUCIONAIS

Dissertação apresentada ao Programa de Pós-Graduação em Engenharia Elétrica da Universidade Federal de Uberlândia, como parte dos requisitos para a obtenção do título de Mestre em Engenharia Elétrica.

Data de defesa: 12 de dezembro de 2025

Comissão Julgadora:

Prof. Dr. Ederson Rosa da Silva
Membro Interno - UFU

Profa. Dra. Milena Bueno Pereira Carneiro
Coorientadora

Prof. Dr. Cláriton Rodrigues Bernadelli
Membro Externo - UFTM

Profa. Dra. Cristiane de Fátima dos Santos Cardoso
Membro Externo - IF Goiano

Prof. Dr. Rafael Augusto da Silva
Membro Interno - UFU

Uberlândia
2025

UNIVERSIDADE FEDERAL DE UBERLÂNDIA
Coordenação do Programa de Pós-Graduação em Engenharia Elétrica
Av. João Naves de Ávila, 2121, Bloco 3N - Bairro Santa Mônica, Uberlândia-MG, CEP 38400-902
Telefone: (34) 3239-4707 - www.posgrad.feelt.ufu.br - copel@ufu.br

ATA DE DEFESA - PÓS-GRADUAÇÃO

Programa de Pós-Graduação em:	Engenharia Elétrica				
Defesa de:	Dissertação de Mestrado, 812, PPGEELT				
Data:	Doze de dezembro de dois mil e vinte e cinco	Hora de início:	8:30	Hora de encerramento:	11:00
Matrícula do Discente:	12322EEL001				
Nome do Discente:	Bianca Bertoldo de Oliveira				
Título do Trabalho:	Detecção de Maturidade Óssea utilizando algoritmo YOLO e Redes Neurais Convolucionais				
Área de concentração:	Processamento da Informação				
Linha de pesquisa:	Processamento Digital de Sinais e Redes de Comunicação				
Projeto de Pesquisa de vinculação:	Coordenador do projeto: Antônio Cláudio Paschoarelli Veiga. Título do projeto: Processamento digital de imagens e vídeos aplicado à sistemas biométricos. Vigência do projeto: em andamento.				

Reuniu-se através de videoconferência, a Banca Examinadora designada pelo Colegiado do Programa de Pós-graduação em Engenharia Elétrica, assim composta:

Doutores: Ederson Rosa da Silva (UFU), Rafael Augusto da Silva (UFU), Cláriton Rodrigues Bernadelli (UFTM), Cristiane de Fátima do Santos Cardoso (IF Goiano) e Milena Bueno Pereira Carneiro, coorientadora da discente.

Iniciando os trabalhos o presidente da mesa, Dr. Ederson Rosa da Silva, apresentou a Comissão Examinadora e a candidata, agradeceu a presença do público, e concedeu à discente a palavra para a exposição do seu trabalho. A duração da apresentação da discente e o tempo de arguição e resposta foram conforme as normas do Programa.

A seguir o senhor presidente concedeu a palavra, pela ordem sucessivamente, aos examinadores, que passaram a arguir a candidata. Ultimada a arguição, que se desenvolveu dentro dos termos regimentais, a Banca, em sessão secreta, atribuiu o resultado final, considerando a candidata:

APROVADA.

Esta defesa faz parte dos requisitos necessários à obtenção do título de Mestre. O competente diploma será expedido após cumprimento dos demais requisitos, conforme as normas do Programa, a legislação pertinente e a regulamentação interna da UFU.

Nada mais havendo a tratar foram encerrados os trabalhos. Foi lavrada a presente ata que após lida e achada conforme, foi assinada pela Banca Examinadora.

Documento assinado eletronicamente por **Milena Bueno Pereira Carneiro**, **Professor(a) do Magistério Superior**, em 12/12/2025, às 15:09, conforme horário oficial de Brasília, com fundamento no art. 6º, § 1º, do Decreto nº 8.539, de 8 de outubro de 2015.

Documento assinado eletronicamente por **Ederson Rosa da Silva**, **Professor(a) do Magistério Superior**, em 12/12/2025, às 15:36, conforme horário oficial de Brasília, com fundamento no art. 6º, § 1º, do Decreto nº 8.539, de 8 de outubro de 2015.

Documento assinado eletronicamente por **Rafael Augusto da Silva**, **Professor(a) do Magistério Superior**, em 12/12/2025, às 15:39, conforme horário oficial de Brasília, com fundamento no art. 6º, § 1º, do Decreto nº 8.539, de 8 de outubro de 2015.

Documento assinado eletronicamente por **Cláriton Rodrigues Berndelli**, **Usuário Externo**, em 15/12/2025, às 08:52, conforme horário oficial de Brasília, com fundamento no art. 6º, § 1º, do Decreto nº 8.539, de 8 de outubro de 2015.

Documento assinado eletronicamente por **Cristiane de Fátima dos Santos Cardoso**, **Usuário Externo**, em 17/12/2025, às 06:28, conforme horário oficial de Brasília, com fundamento no art. 6º, § 1º, do Decreto nº 8.539, de 8 de outubro de 2015.

A autenticidade deste documento pode ser conferida no site https://www.sei.ufu.br/sei/controlador_externo.php?acao=documento_conferir&id_orgao_acesso_externo=0, informando o código verificador **6937384** e o código CRC **85B5CDA7**.

Referência: Processo nº 23117.088469/2025-17 SEI nº 6937384

Dedico este trabalho a minha família e amigos.
Muito obrigada.

AGRADECIMENTOS

Ao meu pai **Durval**, minha madrasta **Cyntia** e meu irmão **Henrick** pelo apoio e incentivo.

Ao meu marido **Gustavo** pelo amor e suporte.

Ao **Prof. Dr. Antonio Cláudio** e à **Profª. Drª. Milena**, pela atenção dedicada ao longo de todo o projeto.

Aos meus familiares e amigos que me apoiaram ao longo dessa jornada.

Ao Programa de Pós-Graduação em Engenharia Elétrica, por possibilitar a realização deste trabalho e por todo o suporte prestado.

À CAPES pelo apoio financeiro a esta pesquisa.

"The good thing about science is that it is
true whether or not you believe in it."
Neil deGrasse Tyson

RESUMO

A avaliação da idade óssea é uma ferramenta diagnóstica essencial na prática pediátrica e endocrinológica, utilizada para monitorar a maturação esquelética e auxiliar no diagnóstico de distúrbios de crescimento. Os métodos tradicionais de avaliação, como os atlas de Greulich and Pyle (GP) ou o sistema de pontuação de Tanner-Whitehouse (TW), dependem da análise visual de radiografias da mão e punho. Contudo, esses métodos manuais são reconhecidamente demorados, subjetivos e suscetíveis a uma variabilidade inter e intraobservador considerável, o que pode impactar a consistência do acompanhamento clínico. O aprendizado de máquina surge como uma solução para reduzir essas limitações, oferecendo o potencial de serem objetivos, rápidos e reprodutíveis. Este trabalho propõe uma metodologia híbrida e automatizada para a avaliação. O fluxo desenvolvido consiste em três etapas principais: primeiramente, as imagens passam por uma fase de pré-processamento, onde a equalização de histograma é aplicada para normalizar o brilho e o contraste. Em seguida, um modelo de detecção de objetos (YOLO) é utilizado para localizar e segmentar automaticamente três regiões de interesse ósseas (articulação, metacarpo e carpo). Por fim, uma arquitetura de regressão customizada é empregada, onde três Redes Neurais Convolucionais (CNNs) independentes processam as imagens. As saídas dessas redes são então fundidas, através de uma média ponderada ajustada, e combinadas com a informação demográfica do sexo do paciente, servindo como entrada para um Perceptron de Múltiplas Camadas (MLP) que realiza a predição final. A avaliação final do modelo, conduzida em um conjunto de teste isolado, resultou em um Erro Médio Absoluto (MAE) de 6,13 meses e um Erro Percentual Absoluto Médio (MAPE) de 6,45%. Estes resultados validam a eficácia da arquitetura híbrida proposta, demonstrando que a fusão de características regionais especializadas e dados demográficos produz um modelo com alta capacidade de generalização e precisão competitiva em relação ao estado da arte, sendo clinicamente relevante para o apoio diagnóstico.

Palavras-chave: aprendizado de máquina. detecção de objetos. idade óssea. python. processamento digital de imagens. redes neurais convolucionais. YOLO.

ABSTRACT

Bone age assessment is an essential diagnostic tool in pediatric and endocrinological practice, used to monitor skeletal maturation and assist in the diagnosis of growth disorders. Traditional assessment methods, such as the Greulich and Pyle (GP) atlas or the Tanner-Whitehouse (TW) scoring system, rely on the visual analysis of hand and wrist radiographs. However, these manual approaches are known to be time-consuming, subjective, and prone to considerable inter- and intra-observer variability, which may compromise the consistency of clinical follow-up. Machine learning emerges as a solution to reduce these limitations, offering the potential to be objective, fast, and reproducible. This work proposes a hybrid and fully automated methodology for bone age evaluation. The developed pipeline consists of three main stages: first, the images undergo a pre-processing phase, in which histogram equalization is applied to normalize brightness and contrast. Next, an object detection model (YOLO) is used to automatically locate and segment three bone regions of interest (joint, metacarpal, and carpal). Finally, a custom regression architecture is employed, in which three independent Convolutional Neural Networks (CNNs) process the images. The outputs of these networks are then fused through an adjusted weighted average and combined with the patient's sex information, serving as input to a Multilayer Perceptron (MLP) that performs the final prediction. The final evaluation of the model, conducted on an isolated test set, resulted in a Mean Absolute Error (MAE) of 6.13 months and a Mean Absolute Percentage Error (MAPE) of 6.45%. These results validate the effectiveness of the proposed hybrid architecture, demonstrating that the integration of specialized regional features with demographic data yields a model with high generalization capability and competitive accuracy compared to the state of the art, making it clinically relevant for diagnostic support.

Keywords: bone age. convolutional neural networks. digital image processing. machine learning. object detection. python. YOLO.

LISTA DE FIGURAS

LISTA DE TABELAS

LISTA DE ABREVIATURAS E SIGLAS

AHE	*Adaptive Histogram Equalization* (Equalização Adaptativa de Histograma)
CNN	*Convolutional Neural Network* (Rede Neural Convolucional)
CNNs	*Convolutional Neural Networks* (Redes Neural Convolucionais)
CLAHE	*Contrast Limited Adaptive Histogram Equalization* (Equalização de Histograma Adaptativa Limitada por Contraste)
EQ	Erro Quadrático
EIA	Escoliose Idiopática do Adolescente
GP	Greulich & Pyle
MLP	*Multilayer Perceptron* (Perceptron de Multicamada)
MLPs	*Multilayer Perceptrons* (Perceptrons de Multicamada)
MAPE	*Mean Absolute Percentage Error* (Erro Percentual Absoluto Médio)
MAE	*Mean Absolute Error* (Erro Absoluto Médio)
MSE	*Mean Squared Error* (Erro Quadrático Médio)
RNA	Rede Neural Artificial
RNAs	Redes Neurais Artificiais
RMSE	*Root Mean Squared Error* (Raiz do Erro Quadrático Médio)
RSNA	*Radiological Society of North America* (Sociedade de Radiologia da América do Norte)
RHPE	*Radiological Hand Pose Estimation* (Estimativa Radiológica da Postura da Mão
TW	Tanner & Whitehouse
YOLO	*You Only Look Once* (Você Só Olha Uma Vez)

SUMÁRIO

1 INTRODUÇÃO

1.1 Considerações Iniciais

A avaliação da idade óssea é uma prática clínica fundamental e rotineira na pediatria, utilizada para monitorar a maturação esquelética de crianças e adolescentes (Satoh, 2015). A discrepância entre a idade óssea de um paciente e sua idade cronológica é um indicador diagnóstico vital para a investigação de uma variedade de condições, incluindo distúrbios de crescimento, desordens endócrinas (como puberdade precoce ou atrasada) e síndromes pediátricas. Na prática clínica, uma idade óssea 20% abaixo ou acima da idade cronológica é considerada anormal (Melmed *et al.*, 2015). A utilidade deste método cessa quando o indivíduo atinge a maturidade esquelética, que é o ponto em que o crescimento ósseo longitudinal termina, geralmente em torno dos 18 ou 19 anos (Malina *et al.*, 2004).

Historicamente, a avaliação da idade óssea é realizada por radiologistas através da inspeção visual de radiografias da mão e punho esquerdos. Dois métodos clássicos, desenvolvidos em meados do século XX, dominam esta prática: o método de Greulich & Pyle (GP) e o método de Tanner & Whitehouse (TW). O método GP (Greulich; Pyle, 1959), se baseia em uma comparação holística da radiografia do paciente com um atlas de imagens de referência padronizadas, selecionando a imagem do atlas que mais se assemelha à do paciente. Por outro lado, o método TW (Tanner *et al.*, 1975), em suas várias iterações (como o TW2 e TW3), adota uma abordagem mais granular e analítica. Neste método, regiões de interesse específicas, como os ossos carpais e as epífises das falanges, são pontuadas individualmente com base em seus estágios de maturação, e a soma desses escores é convertida em uma idade óssea.

O método GP é o padrão dominante e quase exclusivo na prática clínica diária de endocrinologistas e radiologistas. Sua utilização não é apenas uma escolha preferencial baseada na rapidez, mas uma imposição estrutural dos protocolos de saúde pública. O método TW, apesar de sua comprovada superioridade científica e estatística, permanece confinado a nichos acadêmicos, ensaios clínicos farmacêuticos e casos de extrema complexidade dismórfica, devido à sua inviabilidade temporal no fluxo de trabalho moderno (Kim; Yang, 1998).

1.2 Objetivo

Apesar de sua utilidade clínica estabelecida, os métodos manuais são inerentemente dependentes da experiência do radiologista, consomem um tempo clínico valioso e estão sujeitos a uma variabilidade interobservador e intraobservador significativa, o que pode afetar a consistência do diagnóstico (Thodberg *et al.*, 2008). Neste contexto, a evolução do

aprendizado de máquina, e especificamente das redes neurais profundas (*deep learning*), oferece um potencial transformador para a radiologia diagnóstica. Sistemas de *deep learning* têm demonstrado uma capacidade notável na análise de imagens médicas, fornecendo ferramentas objetivas, rápidas e reprodutíveis que podem auxiliar no diagnóstico (Litjens *et al.*, 2017). A automatização dessa tarefa visa, portanto, mitigar as limitações dos métodos manuais, oferecendo uma avaliação consistente, precisa e instantânea.

O presente trabalho propõe uma metodologia híbrida para a avaliação automatizada da idade óssea baseada em aprendizado profundo. A abordagem consiste em um fluxo de três etapas. Primeiramente, as imagens de radiografias da mão passam por um estágio de pré-processamento, no qual são aplicadas técnicas de realce de contraste com o objetivo de reduzir variações de aquisição, melhorar a qualidade visual e enfatizar estruturas ósseas relevantes. Em seguida, um modelo de detecção de objetos *You Only Look Once* (Você Só Olha Uma Vez) (YOLO) é empregado para localizar e segmentar automaticamente múltiplas regiões de interesse da mão. Por fim, essas regiões são processadas por *Convolutional Neural Networks* (Redes Neurais Convolucionais) (CNNs) especializadas em cada região, cujas saídas, combinadas com a informação de sexo do paciente, são integradas por um *Multilayer Perceptron* (Perceptron de Múltiplas Camadas) (MLP) para a regressão final da idade óssea. O objetivo final é desenvolver e validar um sistema de diagnóstico robusto e preciso, capaz de generalizar para novos dados e servir como uma ferramenta de apoio confiável para endocrinologistas pediátricos e radiologistas.

1.3 Hipóteses

Este trabalho parte de hipóteses metodológicas específicas. A hipótese central é que um modelo de *deep learning* que analisa regiões de interesse anatômicas de forma especializada e, em seguida, funde essa informação é capaz de alcançar um desempenho superior ao de modelos que analisam a imagem da mão de forma holística. Assume-se que um detector de objetos moderno (YOLO) pode ser treinado com um conjunto de dados limitado para extrair essas regiões com alta precisão, viabilizando a arquitetura regional.

Adicionalmente, levanta-se a hipótese de que a inclusão de múltiplos conjuntos de dados no treinamento e a integração de informações demográficas são fatores cruciais para melhorar a robustez e a generalização do modelo final. Espera-se que o modelo proposto alcance um resultado competitivo com o estado da arte, validando a eficácia da arquitetura híbrida.

1.4 Estrutura do trabalho

O presente trabalho foi dividido nos seguintes capítulos:

- Capítulo 1 - Introdução - Introdução geral do trabalho, contendo considerações sobre

os métodos manuais existentes, a motivação e objetivos definidos e a estrutura do trabalho.

- Capítulo 2 - Literatura Relacionada - Aborda a literatura relacionada, revisando os métodos clássicos de radiologia, e os algoritmos aplicados ao problema da avaliação da idade óssea, métodos utilizados e resultados obtidos.
- Capítulo 3 - Fundamentação Teórica - Detalha a fundamentação teórica, cobrindo os conceitos de Processamento Digital de Imagens, Redes Neurais Artificiais, hiperparâmetros e as métricas de performance utilizadas.
- Capítulo 4 - Metodologia do Trabalho - Descreve a metodologia empregada, incluindo a definição do banco de dados, as ferramentas de software, o pré-processamento, a estratégia de extração de regiões de interesse e o modelo final de detecção da idade óssea.
- Capítulo 5 - Resultados - Apresenta os resultados obtidos, detalhando o desempenho dos modelos de segmentação de regiões e de detecção da idade óssea, e conduz discussões comparativas com o estado da arte.
- Capítulo 6 - Conclusão - Finaliza o trabalho com as conclusões, sintetizando os resultados, destacando as contribuições da pesquisa e propondo trabalhos futuros.

2 LITERATURA RELACIONADA

2.1 Introdução

O desenvolvimento de técnicas eficazes para avaliar a maturidade óssea é essencial para tomada de decisão clínica e precisão do diagnóstico, contribuindo no tratamento adequado de diversas condições de saúde. Novos métodos para automatizar a avaliação da maturidade óssea foram apresentados nos últimos anos devido ao avanço das técnicas de aprendizado de máquina e visão computacional, reduzindo a necessidade de extensas avaliações manuais e diminuindo a possibilidade de erro humano.

Diversas metodologias tradicionais de processamento de imagens têm sido usadas para resolver o problema com diferentes graus de sucesso. Normalmente, tais abordagens envolvem procedimentos de pré processamento que envolvem a segmentação de regiões de interesse e o treinamento de algoritmos de redes neurais, fazendo uma análise de regressão na saída para obter a idade como resultado final. As principais contribuições na área são revisadas, abordando tanto as técnicas tradicionais de avaliação da maturidade óssea quanto os avanços feitos pelos algoritmos de aprendizado de máquina.

Neste capítulo, primeiramente são abordadas as técnicas radiológicas tradicionais que têm sido empregadas há anos para avaliar a maturidade óssea, juntamente com algumas de suas desvantagens. Em seguida, as estratégias baseadas em aprendizado de máquina são discutidas, enfatizando os algoritmos mais populares, seus usos em diferentes contextos e os resultados alcançados.

2.2 Métodos Clássicos de Avaliação

Os métodos radiológicos clássicos se baseiam principalmente na análise manual de radiografias, com destaque para a análise de regiões específicas, como o punho e a mão (Cavallo *et al.*, 2021). Os dois métodos mais amplamente adotados foram desenvolvidos ao longo do século XX e incluem diretrizes elaboradas por William Walter Greulich e Sarah Idell Pyle (Greulich; Pyle, 1959), e James Tanner e R.H. Whitehouse (Tanner *et al.*, 1975). Cada um deles apresenta diferentes abordagens para avaliação da maturação óssea, variando no detalhamento e na precisão das suas escalas. Em um momento de transição para métodos digitais e automáticos, compreender as bases e limitações desses métodos é essencial para valorizar a evolução da radiologia e os avanços tecnológicos que buscam aprimorar a precisão diagnóstica.

2.2.1 Greulich and Pyle

A publicação de 1959 do Atlas Radiográfico de Desenvolvimento Esquelético da Mão e do Pulso de Greulich e Pyle (GP) continua sendo uma das técnicas mais amplamente utilizadas pelos radiologistas para determinar a idade óssea (Sanctis *et al.*, 2014). Aproximadamente 14.000 radiografias da mão esquerda foram obtidas e cuidadosamente analisadas durante o estudo, que envolveu cerca de 1.000 crianças norte americanas examinadas em intervalos de 3 a 12 meses entre 1931 e 1942. A aparência das superfícies articulares, as correlações de tamanho e a geometria dos ossos nos centros de ossificação são a base para esta abordagem de determinação da idade óssea, com referências separadas para os sexos feminino e masculino. Comparando esses critérios selecionados com centenas de crianças saudáveis de idades similares, eles foram então capazes de estimar os desvios padrão previstos.

Ao comparar as radiografias da mão do paciente com a referência mais próxima listada no atlas, a idade óssea é determinada, expressa em meses. Na versão original da abordagem de Greulich e Pyle, cada osso da mão e do punho é intrinsecamente comparado com radiografias "normais" de referência tiradas em várias idades. A maioria das instituições emprega uma versão "rápida" modificada deste procedimento (Bull *et al.*, 1999), na qual é escolhida a correspondência mais próxima entre a aparência geral de uma determinada radiografia e as radiografias de referência. Este método atualizado pode ser menos preciso que o original, mas é significativamente mais rápido.

Como o método GP se baseia no estudo de uma população específica de pacientes de origem socioeconômica branca de classe média a alta, uma possível desvantagem é sua aplicabilidade a populações diversas modernas. De acordo com pesquisas anteriores realizadas na Ásia, Europa e África, o desenvolvimento esquelético destes jovens não atende às diretrizes de Greulich e Pyle (Zafar *et al.*, 2010), (Büken *et al.*, 2007), (Jiménez-Castellanos *et al.*, 1996), (Kowo-Nyakoko *et al.*, 2023), demonstrando que esse método não seria representativo para todo o espectro étnico de crianças na sociedade.

2.2.2 Tanner and Whitehouse

Desenvolvida numa amostra de cerca de 2.600 crianças britânicas de classes socioeconômicas média e baixa, acompanhadas por intervalos variando de aproximadamente 6 meses a 1 ano, a técnica de Tanner e Whitehouse (TW) foi descrita amplamente pela primeira vez em 1975. Foi posteriormente refinada para TW2 em 1983, utilizando dados adicionais recolhidos na Europa Ocidental e na Ásia. A terceira edição TW3, que incluiu dados da América do Norte e da Europa e pontuações de maturidade modificadas, foi publicada em 2001.

A abordagem considera um exame minucioso das características estruturais das regiões ósseas de interesse. Cada osso recebe uma pontuação determinada pelo sexo e nível

de desenvolvimento do paciente. Desta forma, cada região clinicamente relevante recebe uma pontuação de maturidade, que normalmente é classificada de A até I. Existem até três critérios distintos para cada estágio: a pontuação RUS (que inclui a falange, a ulna e os ossos metacarpais), a pontuação CARPAL (que inclui apenas os ossos do carpo), e as pontuações para 20 ossos específicos do punho e da mão. Cada estágio recebe então um valor numérico. Após somar esses números, obtém-se um escore esquelético, a partir do qual a idade esquelética pode ser obtida diretamente nas tabelas.

O principal benefício deste procedimento é que ele minimiza a variabilidade interoperador ao avaliar cada fragmento ósseo. Embora o método radiográfico TW tenha sido utilizado e aceito por vários grupos étnicos (Zhang *et al.*, 2008), (Shah *et al.*, 2021), (Pinchi *et al.*, 2014), e seja ideal para uso devido à sua capacidade de avaliar centros de ossificação separadamente, radiologistas e endocrinologistas consideram-no menos aceitável devido à sua complexidade na determinação da idade óssea (Buckler, 1983). É necessária muita experiência e mesmo observadores qualificados podem obter resultados diferentes. O posicionamento inadequado das mãos pode alterar significativamente a aparência dos ossos, causando aumentos significativos na pontuação esquelética, assim, mesmo a precisão deste método é limitada. No entanto, é mais objetivo, preciso e reprodutível que o método GP, especialmente para pesquisa (Khan; Elayappen, 2012).

2.3 Algoritmos de Machine Learning

O cálculo computadorizado da idade óssea a partir de radiografias de punho existe há aproximadamente três décadas (Mughal *et al.*, 2014). Geralmente, esses métodos englobam etapas de pré-processamento e segmentação das imagens em regiões de interesse, além do treinamento de algoritmos de Redes Neurais Artificiais (RNAs) de regressão. Para a avaliação dos resultados de predição, a métrica mais comumente utilizada e apresentada em todos os trabalhos analisados foi o Mean Absolute Error (Erro Absoluto Médio) (MAE), que indica a média de desvio das previsões. Quanto menor o valor do MAE, mais preciso é o modelo. Os detalhes matemáticos desse indicador serão discutidos mais profundamente no capítulo 3.

O trabalho de (Chu *et al.*, 2018) foi dividido em duas partes, a primeira para segmentar as imagens, e a segunda para a avaliação da idade óssea. A segmentação foi performada utilizando uma rede neural para geração de máscaras no formato preciso da mão do paciente, baseada em U-Net e VGG16. As imagens resultantes dessa aplicação foram utilizadas para o treinamento do modelo proposto, também baseado no VGG16. O resultado alcançado foi um MAE de 5,98 meses.

(Iglovikov *et al.*, 2018) realizou testes com segmentação da mão inteira e também com extração de regiões de centros de ossificação. O melhor resultado obtido de MAE de 6,10 meses foi encontrado realizando a técnica de *Ensemble Learning*, que consiste em

combinar diversos modelos de predição mais simples e produzir a partir desses um modelo agrupado mais complexo, e para esse método, foram agrupados 15 modelos treinados utilizando arquitetura VGG-Net.

Em (Wibisono *et al.*, 2019), nenhum método de pré-processamento e segmentação foi empregado. O trabalho compõe um estudo de comparação entre duas redes, VGG16 e MobileNet, com resultados de 14,78 meses e 17,09 meses respectivamente.

Em (Pan *et al.*, 2020) o método proposto alcançou MAE de 8,59 meses, onde foi aplicada uma pré etapa de segmentação precisa da mão inteira, performada com modelo U-Net e Deep Active Learning (AL), além de normalização no brilho das imagens. Foram realizados testes com diversos modelos pré-treinados diferentes, e o que obteve o melhor resultado apresentado foi uma combinação de dois modelos Inception-ResNet-V2.

Na pesquisa de (Mehta *et al.*, 2021) o pré-processamento é realizado manipulando uma função que permite desenhar uma região retangular para corte, obtendo manualmente das radiografias originais apenas a região ao entorno da mão, e melhorando a luminância das imagens com correção gama. O método proposto utilizou InceptionV3 atingindo um resultado MAE de 5,92 meses.

O método de (Zulkifley *et al.*, 2021) consiste em uma primeira parte de segmentação da região da mão do fundo com DeepLab V3+, sendo depois girada para uma posição vertical com MobileNet V1. A segunda parte é a predição da idade óssea, realizada com arquitetura Xception remodelada, que atingiu um valor de erro de 7,69 meses.

(Guo *et al.*, 2022) conduziu um estudo mais extenso avaliando modelos com diferentes regiões ósseas segmentadas da mão e punho, separadas e em conjunto. Para a etapa de segmentação, a arquitetura do modelo Inception-ResNet-V2 foi utilizada. O resultado final obtido na etapa de avaliação da idade foi de 6,07 meses, utilizando o modelo híbrido combinando todas as regiões extraídas.

(Andleeb *et al.*, 2025) desenvolveu um algoritmo baseado em DenseNet201, sem performar nenhuma segmentação nas imagens em regiões de interesse, acoplando a saída da rede pré-treinada em uma rede neural customizada para a etapa de regressão e estimativa da previsão. O MAE final foi de 4,87 meses.

Em (Sirati-Amsheh *et al.*, 2025) nenhuma técnica de segmentação é aplicada, utilizando apenas redimensionamento necessário das amostras. A metodologia proposta foi de utilizar uma rede Autoencoder, para extrair features das imagens, utilizando as saídas como entrada da rede neural convolucional padrão, onde obtiveram o resultado final de 9,30 meses.

Todos os estudos apresentados foram resumidos avaliando a aplicação de segmentação em regiões de interesse, se foram geradas novas imagens (*Data Augmentation*) a partir de cópias alteradas, e se houve a reutilização de um modelo pré-treinado (*Transfer*

Learning). Com relação ao número de amostras, todos os métodos propostos utilizaram a base de dados da Radiological Society of North America (Sociedade de Radiologia da América do Norte) (RSNA), que possui 12611 imagens de treino e validação. Em 3 dos 9 trabalhos citados foi empregado *Data Augmentation*, e todos utilizaram *Transfer Learning*. A Tabela 1 ilustra esses resultados:

Tabela 1 – Comparação de resultados para avaliação por RNAs de regressão

Trabalho	**Segmentação**	**Data Augmentation**	**Transfer Learning**	**MAE (meses)**
(Chu *et al.*, 2018)	Mão inteira	Sim	U-Net e VGG16	5,98
(Iglovikov *et al.*, 2018)	Mão inteira e centros de ossificação	Sim	VGG-Net	6,10
(Wibisono *et al.*, 2019)	-	Não	VGG16 e MobileNet	14,78
(Pan *et al.*, 2020)	Mão inteira	Não	U-Net e Inception-ResNet-V2	8,59
(Mehta *et al.*, 2021)	Mão inteira	Sim	InceptionV3	5,92
(Zulkifley *et al.*, 2021)	Mão inteira	Não	DeepLab V3+ e Xception	7,69
(Guo *et al.*, 2022)	6 centros de ossificação	Não	Inception-ResNet-V2	6,07
(Andleeb *et al.*, 2025)	-	Não	DenseNet201	4,87
(Sirati-Amsheh *et al.*, 2025)	-	Não	Autoencoder	9,30

2.4 Considerações Finais

Neste capítulo foram abordadas as principais técnicas utilizadas na avaliação da idade óssea. Na seção 2.2 foram discutidas as principais técnicas fundamentadas em bases teóricas que são empregadas como prática clínica padrão. Por outro lado, a seção 2.3 apresentou as principais técnicas e algoritmos de *machine learning* desenvolvidos nos últimos anos para a previsão computacional e automatizada desse indicador clínico.

Os métodos de segmentação e arquitetura das redes abrangeram uma ampla gama de abordagens, indicando que para o desafio proposto, outras características como robustez do algoritmo e performance em bases de teste podem ser mais importantes para obter melhores resultados.

No capítulo seguinte serão apresentados os fundamentos teóricos de cada método de processamento e dos algoritmos empregados na pesquisa, contemplando suas formulações matemáticas e estatísticas, bem como a descrição do software e da linguagem de programação utilizada.

3 FUNDAMENTAÇÃO TEÓRICA

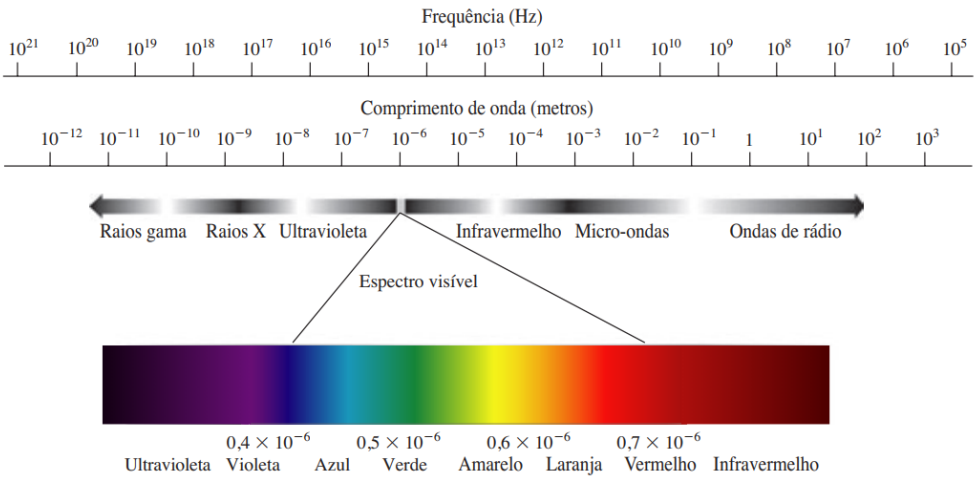
3.1 Processamento Digital de Imagens

A luz é um tipo particular de radiação eletromagnética que pode ser percebida pelo olho humano e apresenta um comportamento ondulatório caracterizado por sua frequência (f), velocidade (c) (2.998 $\times$ 108 m/s) e comprimento de onda (λ). O comprimento de onda é dado em metros, e se relaciona com a frequência por meio da equação (1):

$$\lambda = \frac{c}{f} \tag{3.1}$$

A frequência é expressa em Hertz (Hz), sendo que 1 Hz corresponde a um ciclo completo de uma onda senoidal por segundo. As ondas eletromagnéticas podem ser representadas como ondas senoidais que se propagam no espaço, caracterizadas por seu comprimento de onda. A organização dessas ondas, em função de suas frequências e comprimentos de onda, constitui o espectro eletromagnético, ilustrado na Figura 1.

Figura 1 – Representação do espectro eletromagnético



Fonte: Gonzales e Wintz (1987). Adaptado pela autora.

Dentro desse espectro, a faixa visível ao olho humano corresponde aproximadamente ao intervalo entre 0,43 μm (violeta) e 0,79 μm (vermelho). Nos menores comprimentos de onda, situam-se os raios gama e os raios X. A radiação gama desempenha um papel fundamental em áreas como a medicina, na produção de imagens para diagnóstico, na astronomia, para a observação de fenômenos cósmicos, e em aplicações de monitoramento em ambientes nucleares. Já os raios X, além de sua conhecida utilização em exames médicos, são largamente empregados na odontologia e em inspeções de materiais. A região

do infravermelho já se situa em bandas de comprimentos de onda maiores, cuja principal característica é a emissão de calor, com aplicações voltadas à geração de imagens térmicas, permitindo a detecção de objetos ou seres vivos a partir de suas assinaturas de calor.

Um dispositivo físico com a função de captar a energia emitida em determinada faixa do espectro é capaz de gerar imagens referentes a eventos ou objetos de interesse nessa mesma banda. Após a captura, um digitalizador é responsável por converter os sinais analógicos obtidos pelo sensor em dados digitais, estruturados de modo que possam ser armazenados, processados e posteriormente analisados.

A digitalização de imagens consiste na conversão de um campo de imagem contínua, tal como percebido no mundo real, em uma representação digital capaz de ser manipulada por sistemas computacionais. Esse processo envolve a discretização da imagem em pequenas unidades fundamentais, de tamanho finito, conhecidas como pixels. Cada pixel corresponde a um ponto da imagem e está associado a um valor de intensidade entre 0 e 255 para representar uma cor, e em conjunto, formam um arranjo bidimensional que preserva a estrutura visual do objeto original (Gonzales; Wintz, 1987).

A área de processamento de imagens trabalha com a manipulação desses pixels, com o objetivo de aprimorar a qualidade das informações visuais e extrair informações. Nessa seção serão discutidas as técnicas de pré-processamento utilizadas nesse trabalho, com desenvolvimento da teoria, suas aplicações, propósitos e objetivos.

3.1.1 Canais de cores

Imagens monocromáticas, comumente denominadas imagens em escala de cinza, são formadas a partir da variação de intensidades de um único canal, sem qualquer informação de cor. Cada pixel é associado a um valor numérico que representa sua intensidade luminosa, ou seja, o nível de brilho correspondente àquele ponto da imagem.

Esse processo decorre da quantização, em que o intervalo contínuo de intensidades é discretizado em níveis finitos. Se n for o número de bits de código atribuídos a cada amostra, então o número de níveis de quantização possui amplitude $M = 2^n$. O número de níveis M é escolhido de forma que a qualidade da imagem resultante seja aceitável para os observadores humanos (Acharya; Ray, 2005). Imagens digitais de 8 bits são mais amplamente utilizadas nos processos de visão computacional, e a intensidade pode assumir valores inteiros entre 0 e 255, onde 0 representa o preto absoluto, 255 o branco, e os valores intermediários correspondem aos diferentes tons de cinza que compõem a imagem.

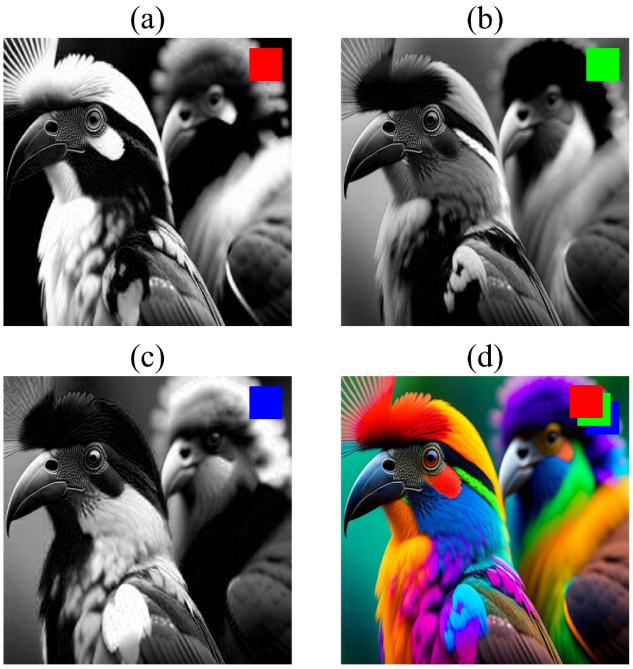
De maneira geral, imagens em escala de cinza apresentam grande utilidade no processamento de imagens, além de serem aplicáveis em diversos outros contextos que demandam menor capacidade computacional em comparação às imagens coloridas. Apesar de a informação cromática ser relevante em muitas situações, quando a cor não é um

fator determinante, as imagens em tons de cinza tornam-se especialmente vantajosas, pois reduzem significativamente o tempo de processamento e o espaço necessário para armazenamento (Tan; Jiang, 2018).

Para a representação de imagens coloridas, foram desenvolvidos modelos de cores padronizados que estabelecem especificações consistentes para uso em diferentes dispositivos. O modelo mais amplamente adotado em aplicações práticas, especialmente em monitores, câmeras e sistemas de vídeo, é o modelo RGB, no qual as cores são formadas pela combinação de três componentes primários: R (*Red* – Vermelho), G (*Green* – Verde) e B (*Blue* – Azul). Cada componente é representado por diferentes níveis de intensidade luminosa, de modo que a mistura dessas intensidades resulta na percepção das diversas cores.

Uma imagem colorida é quantizada separadamente em cada um de seus canais de cor, e cada pixel necessita de n bits por canal, totalizando $3n$ bits por pixel. No caso comum de imagens de 24 bits, são utilizados 8 bits para cada canal, o que resulta em mais de 16,7 milhões de combinações possíveis de cores. Nessa representação, o valor 0 corresponde à ausência de intensidade (preto absoluto), enquanto 255 representa a intensidade máxima da cor em cada canal. Assim, por exemplo, o triplo (0, 255, 0) corresponde ao verde puro, enquanto (255, 255, 255) representa o branco, resultante da soma máxima dos três canais. A Figura 2 demonstra as intensidades de cada canal representadas em níveis de cinza.

Figura 2 – Decomposição de uma imagem colorida em canais de cores segmentados. (a) Canal Vermelho; (b) Verde; (c) Azul. (d) Imagem original



Fonte: Elaborado pela autora.

Os mesmos conceitos definidos para imagens monocromáticas também se aplicam a cada plano de uma imagem colorida. Dessa forma, técnicas de processamento em escala de cinza podem ser utilizadas em cada canal individualmente, permitindo manipulações específicas ou o tratamento independente de R, G e B, antes de serem recombinados para formar a imagem composta.

3.1.2 Equalização de Imagem

Durante o pré-processamento de imagens, é comum a necessidade de corrigir certos elementos que podem comprometer a análise posterior de suas características. Um dos principais fatores é a iluminação, que pode gerar sombras intensas, obscurecendo detalhes de textura e estrutura, ou ainda apresentar variações irregulares que distorcem os resultados obtidos. Uma das técnicas de correção ou aprimoramento de imagens é a equalização de histograma (Pratt, 2007).

O histograma de uma imagem corresponde à representação gráfica da distribuição de intensidades de seus pixels e pode ser entendido como uma estrutura de dados que registra a frequência de ocorrência de cada nível de intensidade presente na imagem, fornecendo uma visão quantitativa sobre sua luminosidade e contraste. O histograma de uma imagem digital com intervalo $[0, L-1]$ é uma função discreta $h(r_k) = n_k$, onde L é o número de níveis de intensidade, r_k é o k-ésimo valor de intensidade e n_k é o número de pixels da imagem com intensidade r_k. A técnica de equalização ajusta a distribuição dos níveis de intensidade da imagem, com o objetivo de criar um histograma com contagens de bin aproximadamente iguais, aproximando-se de uma distribuição em linha reta, melhorando o contraste e resultando, em diversas aplicações, em um aumento significativo no nível de detalhes perceptíveis. A sua função de transformação $T(r)$ é dada pela equação (2):

$$s_k = T(r_k) = \sum_{j=0}^{k} p_r(r_j) = \sum_{j=0}^{k} \frac{n_j}{N}, \quad 0 \leq r_k \leq 1 \quad \text{e} \quad k = 0, 1, ..., L-1 \tag{3.2}$$

onde:

- $T(r_k)$ - Transformação que mapeia um valor de pixel r em um valor de pixel s.
- $p_r(r_j)$ - Função densidade de probabilidade do nível r_j da imagem de entrada.
- n_j - Número de pixels na imagem de entrada que têm nível r_j de tom de cinza.
- N - Número total de pixels na imagem de entrada.

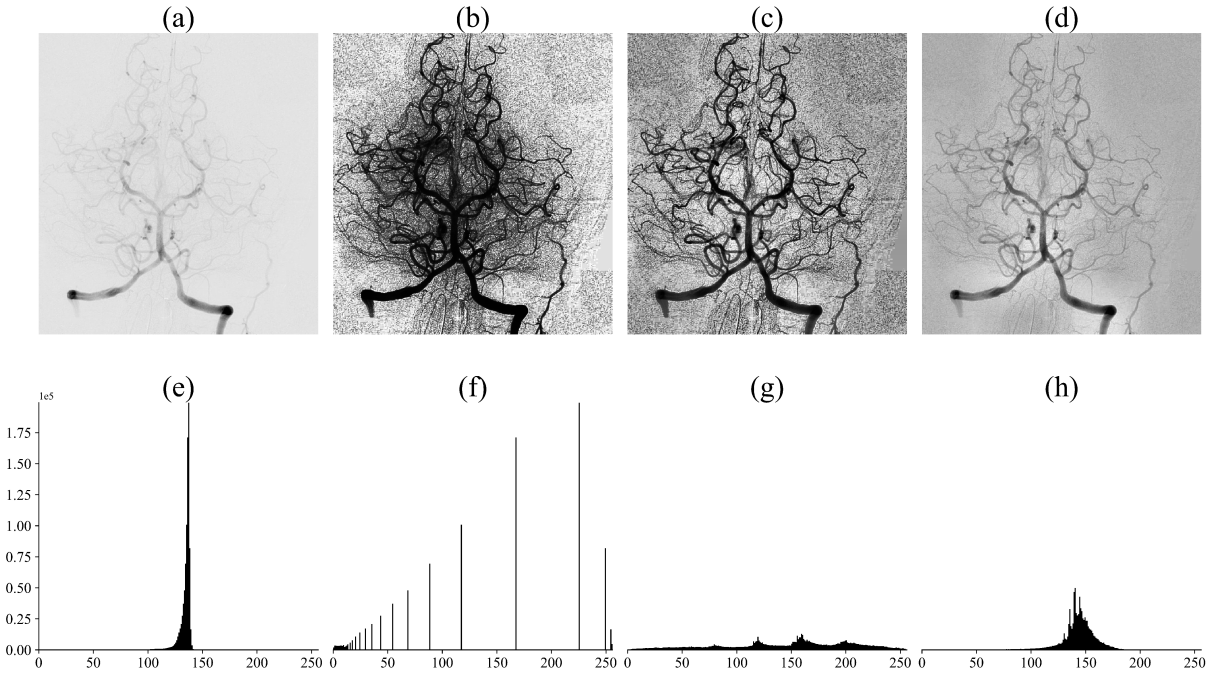
Essa função é um caso específico da classe mais geral de métodos de remapeamento de histogramas, aplicada de maneira global à imagem. Entretanto, essa abordagem apresenta uma limitação importante: ela atua de maneira indiscriminada, podendo realçar o

contraste de ruídos de fundo ao mesmo tempo em que reduz a visibilidade de informações relevantes (Toet; Wu, 2014).

Para contornar esse problema, surgiu o Adaptive Histogram Equalization (Equalização Adaptativa de Histograma) (AHE), , que aplica o processo de equalização de forma local, calculando o histograma em regiões delimitadas por uma janela deslizante bidimensional. Embora mais eficaz em diversos casos, o AHE apresenta a desvantagem de superamplificar o contraste em áreas com baixo nível de variação, resultando em imagens artificialmente exageradas.

Uma solução para essa limitação é o Contrast Limited Adaptive Histogram Equalization (Equalização de Histograma Adaptativa Limitada por Contraste) (CLAHE). Nesse método, o contraste é controlado por meio do recorte do histograma em um valor predefinido antes do cálculo da função de distribuição acumulada. Esse valor limite é ajustado conforme a normalização do histograma ou o tamanho da vizinhança considerada, produzindo resultados mais equilibrados (Shome; Vadali, 2011). A Figura 3 exemplifica o efeito das equalizações em uma captura de angiografia cerebral.

Figura 3 – Equalização em imagem de angiografia cerebral. (a) Imagem original. (b) Equalização global. (c) AHE. (d) CLAHE. (e) a (h) Histograma das imagens (a), (b), (c) e (d) respectivamente



Fonte: Elaborado pela autora.

A imagem original tem pouco contraste entre fundo e vasos, com a maior parte dos pixels concentrados em uma faixa estreita (e). Na equalização global todos os valores de

intensidade são redistribuídos sobre toda a faixa dinâmica (f), gerando uma distribuição quase uniforme mas com grande espaçamento. Os vasos ficam bem visíveis, mas o ruído também é muito amplificado. No histograma do AHE (g) a distribuição de intensidades está espalhada em toda a faixa. Há uma maior variação local e o efeito é uma imagem com contraste muito elevado em regiões pequenas, também exagerando o ruído de fundo. O método CLAHE é mais suave e concentrado (h), sem picos artificiais nem dispersão exagerada. Na imagem, isso se traduz em vasos bem realçados, mas sem amplificar tanto o ruído.

3.1.3 Redimensionamento

O avanço na tecnologia de imagem tornou a geração e a exibição de imagens digitais disponíveis em todos os lugares. Diferentes telas são usadas para visualização, desde monitores de computador de alta resolução até dispositivos móveis de baixa resolução, além da necessidade de impressão de documentos com imagens posicionadas, exigindo frequentemente alterações de tamanho e proporção para a adaptação.

Esse processo de alterar o tamanho de imagens para uma exibição ideal é geralmente denominado como redimensionamento. É uma etapa fundamental no processamento digital e em aplicações de visão computacional. Esse procedimento consiste em reduzir, ampliar ou padronizar o tamanho sem modificar de forma significativa o conteúdo visual relevante. Sua importância está diretamente associada à eficiência computacional, uma vez que imagens de grandes dimensões demandam mais memória e processamento e à qualidade dos resultados obtidos em diferentes contextos.

Em termos de uma imagem bidimensional, significa aumentar/diminuir a largura e a altura calculando os novos valores de pixel. Dado uma imagem I de tamanho $m \times n$ e um tamanho alvo de $m' \times n'$, o objetivo é produzir uma nova imagem I' de tamanho $m' \times n'$ que será um bom representante da imagem I (Vaquero *et al.*, 2010). O método de redimensionamento mais popular é a interpolação dos pixels da imagem original, um processo matemático que utiliza valores conhecidos para estimar valores em pontos desconhecidos. Os métodos mais comuns são: vizinhos mais próximos, bilinear e bicúbica. Contudo, é importante destacar que tais métodos quando mal aplicados podem introduzir distorções perceptíveis, como serrilhados ou borramentos, principalmente quando a proporção da imagem original difere grandemente da proporção da imagem resultante (Dighe; Guru, 2014).

A interpolação do vizinho mais próximo é o método mais simples e computacionalmente mais rápido entre os algoritmos de interpolação. Sua estratégia consiste em atribuir ao novo pixel o valor do pixel mais próximo da posição correspondente na imagem original, obtido pelo arredondamento das coordenadas. Apesar da simplicidade, essa abordagem tende a introduzir artefatos indesejáveis, e por essa razão, seu uso prático é bastante

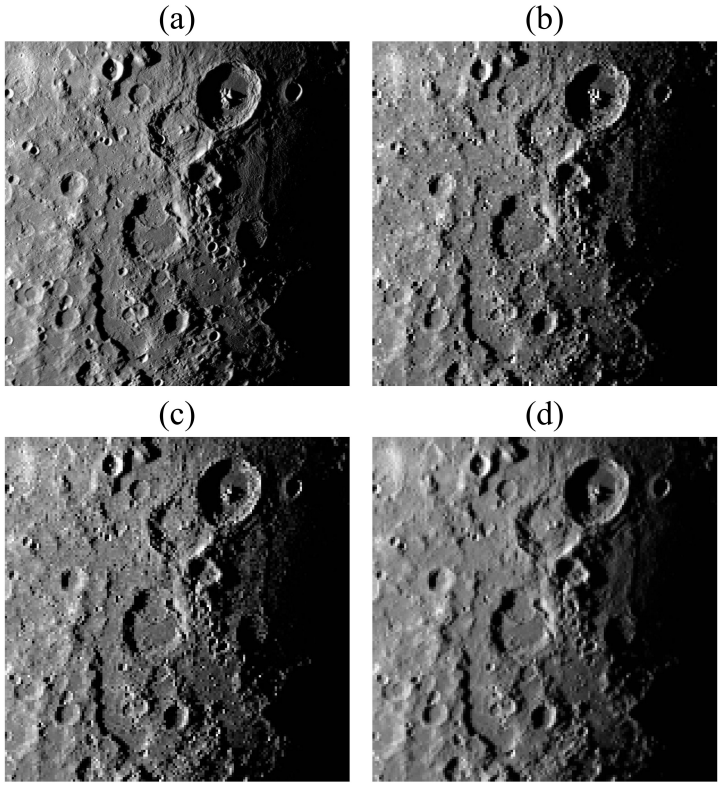
limitado (Parsania *et al.*, 2014).

A interpolação bilinear, por sua vez, utiliza os quatro pixels vizinhos mais próximos para estimar a intensidade de cada ponto da imagem resultante. Embora demande um pouco mais de processamento em comparação ao método anterior, produz resultados visualmente mais agradáveis com transições mais suaves entre pixels.

Já a interpolação bicúbica representa um avanço em relação às anteriores, pois considera um conjunto de 16 pixels vizinhos mais próximos. Esse aumento de complexidade computacional permite preservar detalhes mais finos da imagem e reduzir a ocorrência de distorções, tornando-a superior à bilinear em termos de qualidade visual. Por esse motivo, a interpolação bicúbica é amplamente adotada como padrão em softwares profissionais de edição e processamento de imagens.

A Figura 4 mostra o exemplo de uma imagem de 760 x 760 pixels em (a), redimensionada utilizando os três métodos para 20% de seu tamanho original em (b) até (d), resultando no tamanho final de 152 x 152 pixels. Essa versão reduzida foi ampliada na figura apenas para fins de visualização, de modo a evidenciar de forma clara as diferenças na qualidade final.

Figura 4 – (a) Imagem original. (b) Redimensionamento por interpolação do vizinho mais próximo; (c) Bilinear; (d) Bicúbica



Fonte: Elaborado pela autora.

3.1.4 Data Augmentation

A quantidade e a qualidade dos dados são aspectos que exercem influência direta sobre o desempenho de modelos de aprendizado de máquina. Quando treinados com conjuntos de dados reduzidos, ou expostos a amostras desbalanceadas, a capacidade de generalização do modelo é prejudicada, apresentando desempenho significativamente inferior em dados de teste. Esse cenário é recorrente em aplicações práticas, nas quais a obtenção de dados é frequentemente limitada ou onerosa, e o processo demanda elevado esforço e recursos.

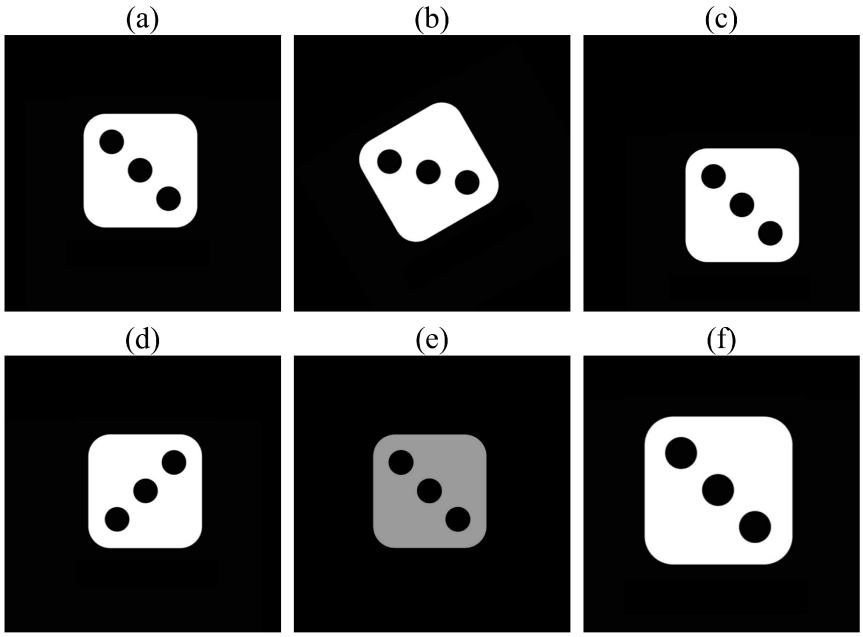
O *Data Augmentation* (Acréscimo de Dados) surge como um conjunto de técnicas voltadas à geração de amostras artificiais de alta qualidade a partir da manipulação de dados já existentes. O princípio fundamental consiste em ampliar artificialmente o conjunto de treinamento por meio da criação de cópias transformadas das amostras originais (Wang *et al.*, 2024). Além de aumentar a diversidade dos dados, essa estratégia contribui para reduzir a discrepância entre os conjuntos de treinamento e as condições encontradas em aplicações do mundo real.

Nesta seção, serão abordados os métodos mais frequentemente usados e de implementação relativamente simples, baseados em transformações espaciais geométricas. Esse tipo de manipulação atua sobre a disposição espacial dos pixels em uma imagem, preservando seus valores originais, de modo a gerar novas amostras a partir da reorganização estrutural do conteúdo visual.

Rotação consiste em girar a imagem em torno de seu centro em diferentes ângulos. Essa transformação permite que o modelo se torne mais invariante a mudanças de orientação, uma característica relevante em tarefas em que a posição relativa do objeto de interesse pode variar. **Translação** refere-se ao deslocamento da imagem ao longo dos eixos horizontal e/ou vertical. Essa técnica auxilia o modelo a lidar com situações em que o objeto não se encontra centralizado ou perfeitamente alinhado no campo de visão. O **espelhamento** aplica uma inversão horizontal ou vertical da imagem, criando uma versão especular da amostra original. Essa operação é particularmente útil em cenários em que a simetria é uma propriedade relevante, evitando que o modelo memorize posições fixas. **Ajuste de brilho** altera a intensidade luminosa da imagem, simulando diferentes condições de iluminação. Dessa forma, o modelo é exposto a variações de contraste e iluminação, tornando-se mais resiliente a essas condições no ambiente real. Por fim, o **zoom** amplia ou reduz a região de interesse da imagem, modificando o nível de proximidade em relação ao objeto.

Apesar da simplicidade dessas técnicas, diversos trabalhos demonstraram sua alta eficácia em diversas tarefas de visão computacional (Mumuni; Mumuni, 2022). A Figura 5 exibe uma imagem original seguida dos resultados das cinco operações para exemplificação.

Figura 5 – Exemplos da aplicação de Data Augmentation. a) Imagem original. b) Rotação de 30º. c) Translação. d) Espelhamento. e) Ajuste de brilho. f) Zoom



Fonte: Elaborado pela autora.

3.2 Anatomia óssea da mão

A mão humana é composta por um conjunto de ossos que desempenham funções fundamentais tanto para a mobilidade quanto para a precisão dos movimentos. Estruturalmente, ela é formada por três grupos principais: carpos, metacarpos e falanges. As falanges, que totalizam 14 em cada mão, são responsáveis pela estrutura dos dedos, dividindo-se em proximais, médias e distais, com exceção do polegar, que possui apenas duas. Os metacarpos, em número de cinco, conectam o carpo às falanges e constituem o esqueleto da palma. Já os ossos do carpo, localizados no punho, são oito pequenos ossos dispostos em duas fileiras que atuam como base de sustentação. Além dessas estruturas, o punho também é composto do rádio e da ulna (Maw *et al.*, 2016).

A idade óssea é estimada a partir da análise da presença, do tamanho, da forma e do grau de fusão dos centros de ossificação. Desde o período neonatal até o final da adolescência, diferentes centros de ossificação surgem e se consolidam em sequência cronológica relativamente previsível. Em crianças mais novas, predominam estruturas cartilaginosas e centros de ossificação ainda pouco desenvolvidos. Com o avanço da idade, esses centros tornam-se progressivamente mineralizados, até a completa fusão, marcando o fim do crescimento longitudinal dos ossos (Gilsanz; Ratib, 2005). Uma representação da anatomia óssea se encontra na Figura 6.

Figura 6 – Anatomia esquelética da mão

Fonte: Elaborado pela autora.

3.3 RNA - Redes Neurais Artificiais

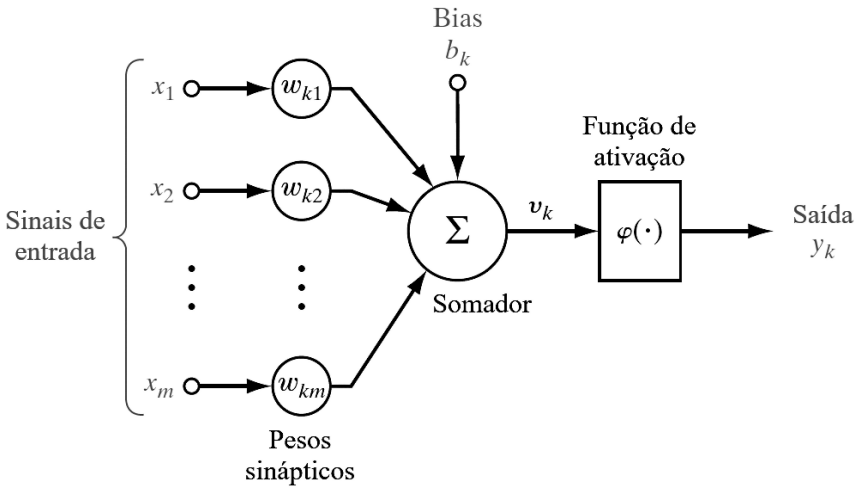
As RNAs são modelos computacionais de autoaprendizagem capazes de realizar predição e processamento de padrões não lineares. Inspiradas em estudos modernos de neurociência, essas redes simulam, de forma simplificada, o funcionamento das conexões neurais do cérebro humano. Sua estrutura é organizada em três componentes principais: a camada de entrada, responsável por receber os dados; uma ou mais camadas ocultas, onde ocorre o processamento interno das informações; e a camada de saída, que fornece o resultado final. As camadas ocultas são formadas por unidades de processamento denominadas neurônios, cada uma conectada a todos os nós da camada anterior. Os neurônios desempenham um papel central na operação da rede, pois são responsáveis por transformar os sinais de entrada em representações úteis na saída (Haykin, 2009).

O modelo de um neurônio k é representado no diagrama de blocos da Figura 7. Define-se três características fundamentais do modelo neural:

- 1. Um grupo de sinapses ou elos de conexão, cada um com seu próprio peso ou força. Em particular, o peso sináptico w_{kj} é multiplicado por um sinal x_j na entrada da sinapse j conectada ao neurônio k.
- 2. Um somador para adicionar os sinais de entrada, que são ponderados pelas forças sinápticas do neurônio; os processos descritos aqui criam um combinador linear.
- 3. Uma função de ativação para restringir a magnitude, ou faixa de amplitude, da

saída de um neurônio.

Figura 7 – Modelo não linear de um neurônio k



Fonte: Haykin (2009). Adaptado pela autora.

Em termos matemáticos, podemos descrever o neurônio k representado na Figura 6 escrevendo as equações (3), (4) e (5).

$$u_k = \sum_{j=1}^{m} w_{kj} x_j \tag{3.3}$$

$$y_k = \varphi(u_k + b_k) \tag{3.4}$$

$$v_k = u_k + b_k \tag{3.5}$$

onde:

- x_j - Sinais de entrada.
- w_{kj} - Pesos sinápticos do neurônio.
- u_k - Saída do combinador linear.
- b_k - *Bias* aplicado a um neurônio.
- $\varphi(\cdot)$ - Função de ativação.
- v_k - Potencial de ativação do neurônio k.
- y_k - Sinal de saída do neurônio.

O *bias* é um elemento que serve para aumentar o grau de liberdade dos ajustes dos pesos, de forma a transladar a função de ativação no eixo.

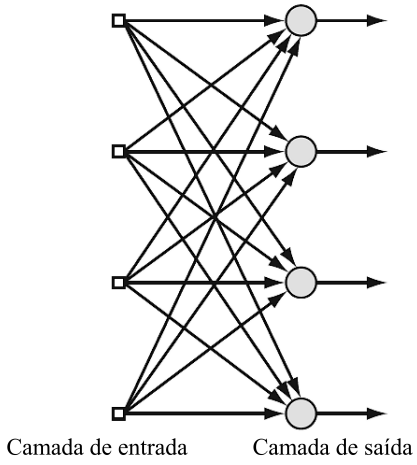
O processo de aprendizagem em redes neurais é conduzido por um algoritmo de aprendizagem, cuja função principal é ajustar os pesos sinápticos da rede de forma estruturada, de modo a atingir o objetivo estabelecido no projeto. A meta é definir uma regra de treinamento que possibilite a atualização dos pesos em cada camada, buscando minimizar a função de erro. Um dos grandes diferenciais desses algoritmos é a capacidade de generalização, ou seja, a habilidade da rede em gerar respostas adequadas mesmo para entradas que não foram apresentadas durante a fase de treinamento. As estruturas de RNA ganharam destaque como a principal metodologia para análise de imagens, devido às suas grandes capacidades de aprendizado e vantagens em lidar com padrões complicados.

A disposição dos neurônios de uma rede neural apresenta estreita relação com o método de aprendizado empregado em seu treinamento. De modo geral, as arquiteturas de redes neurais podem ser agrupadas em três categorias diferentes:

Redes *feedforward* de camada única: Nesse tipo de arquitetura, os neurônios estão organizados de forma que a camada de entrada, composta por nós de origem, projeta-se diretamente sobre a camada de saída, sem a existência de retroalimentação entre elas. Trata-se, portanto, de uma rede *feedforward*, ou direta. A Figura 8 ilustra um exemplo em que quatro nós da camada de entrada se conectam diretamente a quatro nós da camada de saída.

Essa estrutura recebe a denominação de rede de camada única, sendo o termo "camada única" atribuído exclusivamente à camada de saída, onde de fato ocorre o processamento computacional. A camada de entrada, por sua vez, não é considerada nesse contexto, uma vez que não realiza operações de cálculo, funcionando apenas como ponto de fornecimento dos dados à rede.

Figura 8 – Rede *feedforward* com uma única camada de neurônios

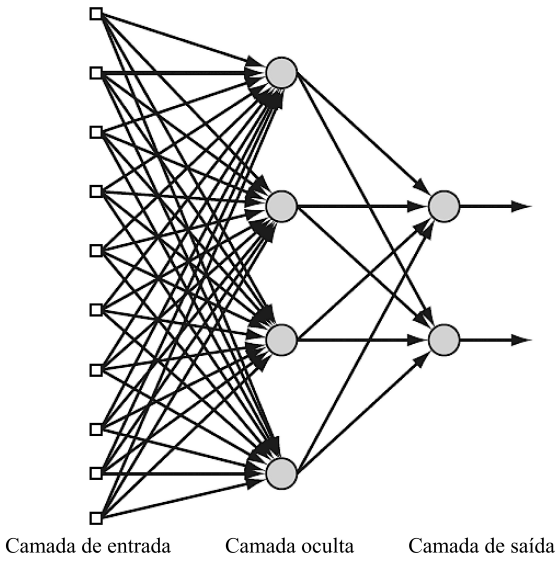


Fonte: Haykin (2009). Adaptado pela autora.

Redes *Feedforward* Multicamadas: Esse tipo de arquitetura é caracterizado pela presença de uma ou mais camadas ocultas, denominadas assim porque não estão diretamente acessíveis nem a partir da entrada nem da saída da rede. A função essencial dessas camadas intermediárias é atuar como um elo entre os dados de entrada e a saída final, possibilitando à rede a capacidade de extrair estatísticas de ordem superior e representar relações mais complexas dos dados.

O processamento ocorre de maneira sequencial: os nós de origem fornecem os elementos que compõem o padrão de ativação inicial, o qual é transmitido para os neurônios da segunda camada. Em seguida, os sinais resultantes dessa camada são propagados para a camada seguinte e assim sucessivamente, até alcançar a camada de saída. O conjunto de ativações produzidas pelos neurônios dessa última camada constitui a resposta final da rede ao padrão de entrada apresentado. Um exemplo de uma rede com arquitetura composta por 10 nós de entrada, 4 neurônios ocultos e 2 neurônios de saída pode ser observado na Figura 9.

Figura 9 – Rede *feedforward* totalmente conectada com uma camada oculta e uma camada de saída

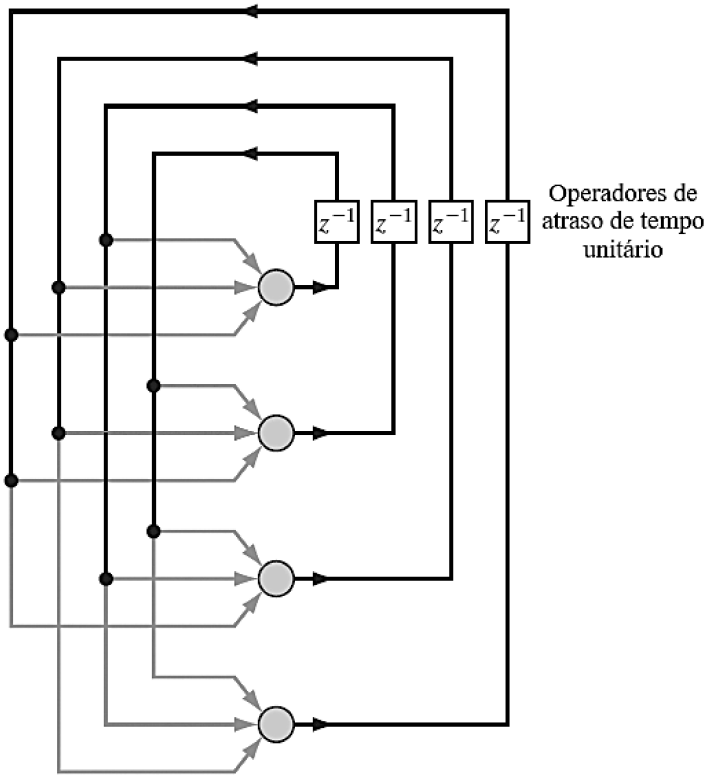


Fonte: Haykin (2009). Adaptado pela autora.

Redes Recorrentes: Diferentemente das redes *feedforward*, onde o fluxo de informação ocorre apenas da entrada até a saída, essa rede caracteriza-se pela presença de pelo menos um *loop* de *feedback*, que incorporam elementos de atraso unitário no tempo (denotados por z^{-1}), armazenando temporariamente os sinais e os reapresentam à rede no próximo instante. Isso significa que a saída de um neurônio em determinado instante pode ser realimentada como entrada em passos de tempo subsequentes, possibilitando que a rede retenha informações do passado e, assim, modele dependências temporais.

Essa rede pode ser composta por apenas uma camada de neurônios, conforme mostrado no gráfico arquitetônico da Figura 10, em que cada neurônio envia sua saída de volta às entradas de todos os outros neurônios (estrutura recorrente totalmente conectada), ou ainda por arquiteturas mais complexas que incluem camadas ocultas adicionais.

Figura 10 – Rede recorrente sem camadas ocultas



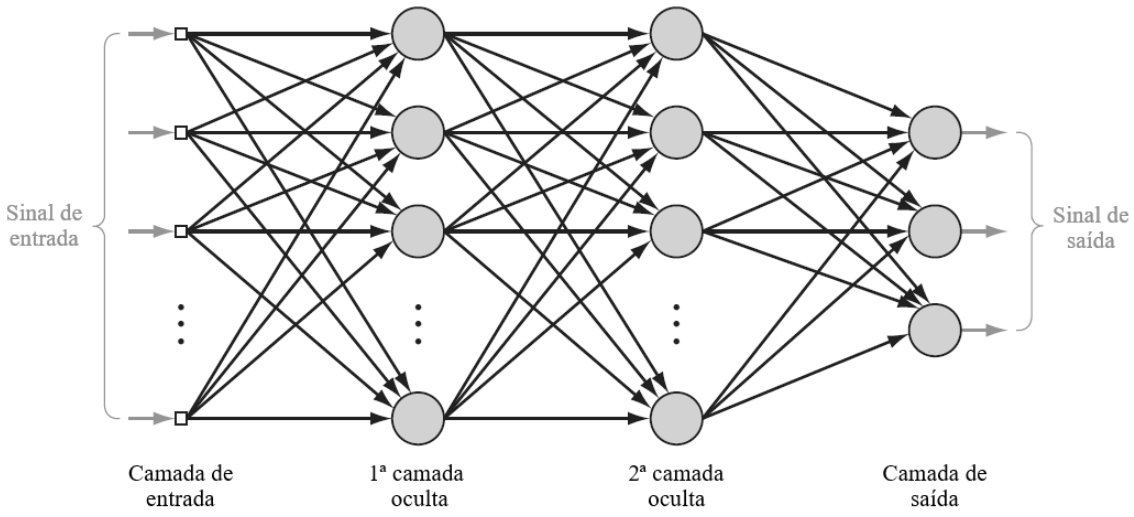
Fonte: Haykin (2009). Adaptado pela autora.

3.3.1 MLP - Multilayer Perceptron

A primeira rede neural descrita algoritmicamente foi o *perceptron*, um modelo de classificação binária proposto por Frank Rosenblatt em 1958. O *perceptron* é o tipo mais básico de rede neural utilizada para categorização de padrões linearmente separáveis, sendo capaz de realizar classificações apenas entre duas classes. A solução para essa limitação veio com a introdução de uma ou mais camadas ocultas de neurônios entre a camada de entrada e a de saída, dando origem ao *Multilayer Perceptron* (Perceptron de Multicamada) (MLP). Ao adicionar essas camadas intermediárias, a rede adquire a capacidade de aprender representações internas e não-lineares dos dados. Essencialmente, as camadas ocultas transformam o espaço de características original em um novo espaço, de maior dimensionalidade, onde o problema se torna linearmente separável para a camada de saída (Pan, 2024).

O MLP pertence a uma classe mais ampla de arquiteturas, sendo um subtipo específico de Redes *Feedforward* Multicamadas. A Figura 11 mostra o gráfico arquitetônico de um MLP com duas camadas ocultas e uma camada de saída.

Figura 11 – Gráfico arquitetônico de um MLP com duas camadas ocultas



Fonte: Haykin (2009). Adaptado pela autora.

O treinamento de uma rede neural é fundamentalmente um problema de otimização, onde objetivo é encontrar o conjunto de parâmetros que minimiza uma função de perda que quantifica a discrepância entre as predições da rede e os valores verdadeiros para um dado conjunto de treinamento. Importante notar que esses processos são aplicáveis não apenas para o MLP, mas também às redes neurais em geral. O método mais comum para realizar essa otimização é o gradiente descendente, um algoritmo iterativo que ajusta os parâmetros na direção oposta ao gradiente da função de perda com respeito aos parâmetros. Os ajustes sucessivos aplicados ao vetor de peso w são na direção da descida mais íngreme, isto é, em uma direção oposta ao vetor do gradiente $\nabla\xi(w)$, que também pode ser resumido pela equação (6):

$$g = \nabla\xi(w) \tag{3.6}$$

Correspondentemente, o algoritmo da descida mais íngreme é descrito formalmente pela equação (7):

$$w(n+1) = w(n) - \eta g(n) \tag{3.7}$$

onde η é uma constante positiva chamada de tamanho do passo ou parâmetro de taxa de aprendizagem (que será melhor descrita na seção 3.2.4), e $g(n)$ é o vetor do

gradiente calculado no ponto $w(n)$. Assim, a regra de atualização para cada parâmetro é dada pela equação (8):

$$\Delta w(n) = w(n+1) - w(n) = -\eta g(n) \tag{3.8}$$

O algoritmo de retropropagação é um método notavelmente eficiente para calcular o gradiente, e funciona determinando sistematicamente como cada peso e *bias* na rede contribui para o erro total. Ele consiste em duas fases, sendo a primeira *forward pass* (passe direto), onde um lote de dados de entrada é propagado através da rede, da primeira à última camada, para gerar as saídas. A função de perda é então usada para calcular o erro entre as saídas previstas e os alvos verdadeiros. A segunda é o *backward pass* (passe reverso), onde o gradiente do erro é calculado primeiro com respeito às ativações da camada de saída e é propagado para trás, camada por camada. Em cada camada, é calculado o gradiente da perda com respeito aos seus parâmetros, e o gradiente da perda com respeito às ativações da camada anterior. Este último é então passado para a camada anterior, de forma sucessiva, até que os gradientes para todos os parâmetros tenham sido calculados (Amari, 1993).

3.3.2 CNN - Convolutional Neural Network

Convolutional Neural Networks (Redes Neural Convolucionais) (CNNs) são baseadas em MLPs projetado especificamente para reconhecer e avaliar formas bidimensionais. Essas redes empregam uma arquitetura única que é adequada para a classificação de imagens, e se valem de três conceitos fundamentais: convolução, pesos e *bias* compartilhados e *pooling*.

Convolução: É o processo central em uma CNN e serve para extrair características relevantes de uma imagem ou sinal, permitindo que a rede detecte padrões como bordas, texturas ou formas específicas. Uma pequena matriz de pesos é definida, chamada comumente de filtro, percorrendo a imagem por um processo de deslizamento com um passo determinado (Nielsen, 2015). Em cada posição, realiza-se o produto ponto a ponto entre os valores da imagem e os valores do filtro, em seguida todos esses produtos são somados, produzindo um único valor escalar, que irá compor a posição equivalente no *feature map* (mapa de características).

A dimensão D_{out} do *feature map* resultante de uma operação de convolução é dada pela equação (9):

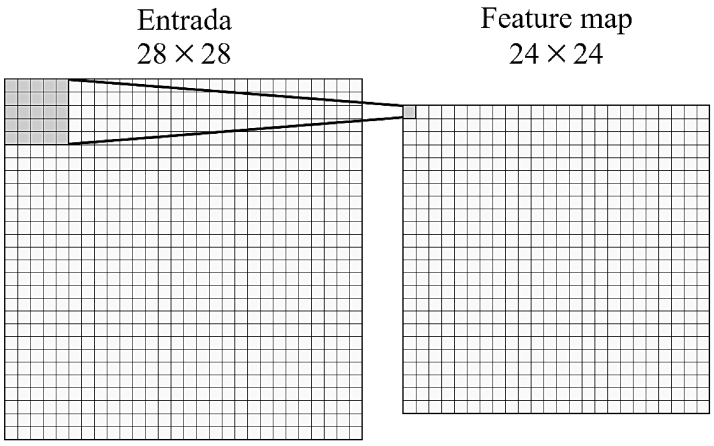
$$D_{out} = \frac{D_{in} - F + 2P}{S} + 1 \tag{3.9}$$

onde:

D_{in} - Dimensão de entrada (altura ou largura).
F - Tamanho do filtro.
P - *Padding* (preenchimento) aplicado na borda da imagem.
S - *Stride* (passo) do filtro.

A Figura 12 mostra uma imagem de tamanho 28x28, com um filtro deslizante de 5x5 com passo 1 no canto superior esquerdo, produzindo uma saída de tamanho 24x24:

Figura 12 – *Feature map* criado a partir da convolução da imagem de entrada



Fonte: Elaborado pela autora.

Pesos e *bias* compartilhados: Os pesos determinam a importância de cada pixel dentro do filtro para detectar padrões específicos, como bordas ou texturas, e o *bias* é um valor adicional que permite ao neurônio ajustar a ativação independentemente dos valores dos pixels, aumentando a flexibilidade do modelo. Para um filtro específico, os mesmos pesos são usados em todas as posições da imagem, aplicando a mesma operação de convolução em cada região, identificando a mesma característica. Cada filtro também possui um único *bias*, que é adicionado a todos os valores do *feature map* gerado pelo filtro. Esse compartilhamento garante eficiência e consistência nas operações.

Camadas de *pooling*: Etapa normalmente empregada imediatamente após à convolução, com o objetivo de reduzir a dimensão espacial do *feature map*, mantendo as informações mais importantes. Uma janela, geralmente de tamanho 2x2 ou 3x3, percorre o *feature map* gerado em passos definidos e substitui os pixels da região por um único valor. Os tipos mais comuns são: ***Max Pooling***, que seleciona o valor máximo, destacando as características mais expressivas, como bordas e texturas fortes; ***Average Pooling*** calcula a média dos valores da janela produzindo uma versão suavizada do feature map; e

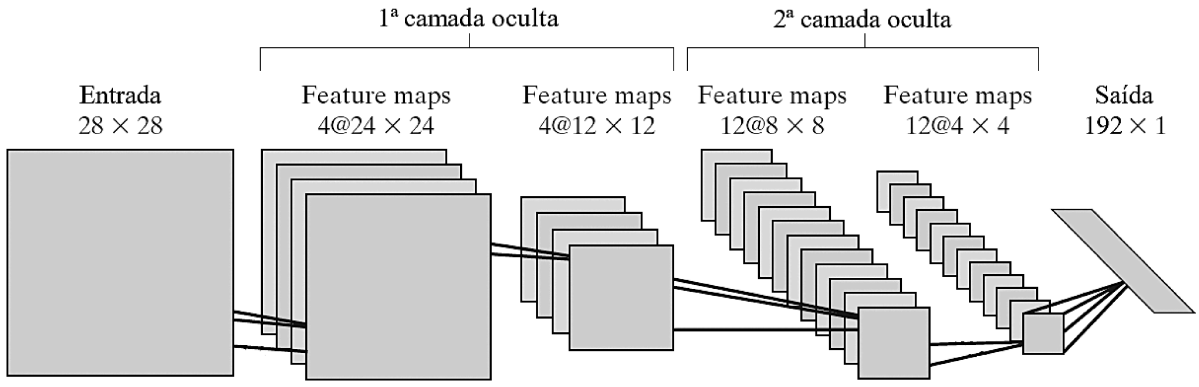
o ***Min Pooling*** que seleciona o valor mínimo, destacando as características mais suaves. Essa técnica diminui a complexidade computacional e introduz invariância a pequenas mudanças na posição de características (Hou *et al.*, 2021).

Cada unidade na camada de *pooling*, por exemplo, pode resumir uma região de 2 × 2 neurônios na camada anterior. Depois de agrupar os 24 x 24 neurônios produzidos a partir da camada convolucional ilustrada no exemplo, temos 12 x 12 neurônios. O agrupamento é aplicado separadamente à cada *feature map*.

Como normalmente é necessário detectar múltiplas características para construir um modelo eficaz, geralmente haverá múltiplos *feature maps* diferentes em cada camada convolucional. A Figura 13 representa a arquitetura de uma CNN, que consiste em uma camada de entrada, duas camadas ocultas e uma camada de saída. A camada de entrada recebe imagens de dimensão 28x28 pixels. Na primeira camada oculta são aplicados 4 filtros convolucionais de tamanho 5x5, produzindo 4 *feature maps* de tamanho 24x24. Cada mapa 24x24 passa por *pooling* de tamanho 2x2, reduzindo suas dimensões pela metade.

Na segunda camada oculta, são aplicados 12 filtros convolucionais (de tamanho 5x5) sobre os 4 mapas anteriores, seguido novamente por um *pooling* de 2x2, obtendo uma saída de tamanho 4x4. Cada mapa 4x4 tem 16 elementos, e multiplicando pelo número de *feature maps* resultantes, têm-se 192 neurônios de entrada para a camada densa. Esses valores são achatados e conectados a uma camada totalmente conectada. A depender do problema proposto, as características dos neurônios resultantes podem ser mapeadas para o número de classes discretas à ser detectada, ou para um único neurônio, caso a saída desejada seja um valor contínuo.

Figura 13 – Rede convolucional para processamento de imagem



Fonte: Haykin (2009). Adaptado pela autora.

3.3.3 YOLO - You Only Look Once

Reconhecimento de imagens é uma área da visão computacional e do aprendizado de máquina que tem como objetivo identificar e classificar objetos, padrões, pessoas, lugares ou qualquer outro elemento presente em uma imagem digital. Um conjunto de objetos similares, com uma ou mais características semelhantes, é considerado como pertencente à mesma classe de padrões (Queiroz; Gomes, 2006).

A tarefa de reconhecimento pode ser divida em 3 partes: **classificação**, que se refere à atribuição de uma categoria; **detecção**, que envolve localizar e rotular múltiplos objetos na mesma imagem; e **segmentação**, que separa cada objeto ou região da imagem pixel a pixel. A detecção de objetos surgiu como um componente crítico em inúmeras aplicações, abrangendo vários campos, como segurança, saúde, veículos autônomos, robótica, e realidade aumentada.

Os modelos que conduzem essas análises podem ser categorizados em dois tipos: detectores de um estágio e detectores de dois estágios. Os detectores de dois estágios dividem a tarefa de reconhecimento em duas partes. A primeira, gera propostas de regiões, que são áreas da imagem onde provavelmente existe um objeto. Para cada região proposta o modelo extrai as características visuais correspondentes e classifica o que tem dentro daquela região. Devido à essa natureza de duas etapas executadas sucessivamente para várias regiões da imagem, e à complexidade do CNN utilizado, os custos computacionais são altos, limitando seu uso em detecções offline e não em cenários de casos em tempo real (Baldovino *et al.*, 2024).

Por outro lado, detectores de estágio único são compostos apenas por uma CNN de alimentação única, que fornece diretamente as caixas delimitadoras dos objetos, bem como suas respectivas classificações. As mudanças oferecidas por esses modelos resultaram em custos computacionais mais baixos, proporcionando maior precisão, e ainda assim mantendo uma alta velocidade de processamento.

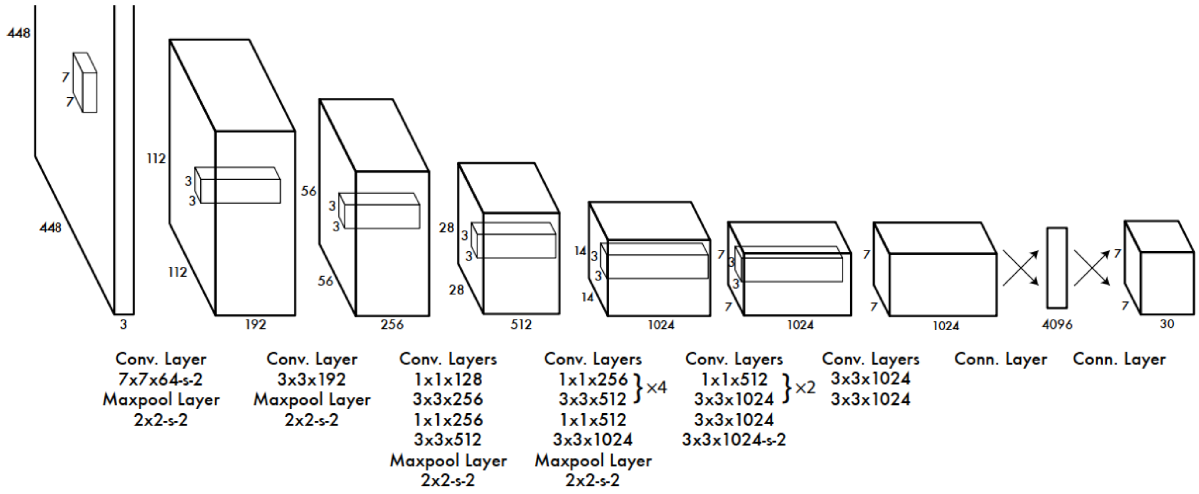
Entre os diferentes algoritmos de detecção de objetos de estágio único, o framework *You Only Look Once* (Você Só Olha Uma Vez) (YOLO), de Joseph Redmon et al., foi publicado na IEEE Conference on Computer Vision and Pattern Recognition (CVPR) de 2016 (Redmon *et al.*, 2016). Ele apresentou pela primeira vez uma abordagem ponta a ponta em tempo real para reconhecimento de objetos. Seu nome se refere ao fato de que essa arquitetura é capaz de realizar a tarefa de detecção e classificação com uma única passagem pela rede.

O YOLO analisa globalmente a imagem ao fazer previsões. Ao contrário das técnicas baseadas em janela deslizante e proposta de região, esse modelo analisa a imagem inteira durante o treinamento e o teste, codificando implicitamente informações contextuais sobre as classes, bem como sua aparência, aprendendo representações generalizáveis de

objetos (Terven *et al.*, 2023). Desde o seu início, a família YOLO evoluiu por meio de múltiplas iterações, cada uma delas baseada nas versões anteriores para abordar limitações e aprimorar o desempenho. A versão lançada em 2024, o YOLOv11, de autoria de Glenn Jocher e Jing Qiu (Jocher; Qiu, 2024), é o modelo mais recente e avançado da série YOLO da empresa Ultralytics.

Na arquitetura do YOLO, as camadas convolucionais profundas não apenas extraem características, mas são organizadas para produzir um mapa de características espacialmente alinhado à imagem, no qual cada posição do mapa corresponde a uma região fixa da imagem de entrada. À medida que a rede reduz a resolução espacial por meio de convoluções e pooling, ela aumenta a abstração semântica dessas regiões, de modo que cada célula do mapa final codifica o que existe naquela região e onde está. As últimas camadas convolucionais atuam com regressão densa, transformando esse mapa em um tensor de saída no qual, para cada célula da grade, são previstos simultaneamente as coordenadas e dimensões das caixas delimitadoras e as probabilidades de classe. A arquitetura base do modelo pode ser visualizada na Figura 14.

Figura 14 – Arquitetura do YOLO



Fonte: Redmon *et al.* (2016).

3.3.4 Hiperparâmetros e Configurações

Os hiperparâmetros e as configurações estruturais de uma rede neural são variáveis que influenciam a topologia da rede e são fundamentais para o sucesso do processo de treinamento e determinação da sua capacidade de generalização. Diferentemente dos parâmetros aprendidos e ajustados durante o treinamento (os pesos sinápticos), esses elementos precisam ser estabelecidos previamente e exercem influência determinante sobre o comportamento de aprendizado. Uma seleção adequada permite a obtenção de redes com melhor desempenho, velocidade de convergência otimizada e capacidade de generalização

robusta, além de reduzir o risco de *overfitting* e o custo computacional (Raiaan *et al.*, 2024).

a) Classificação e Regressão

No contexto de aprendizado de máquina supervisionado, os modelos são treinados a partir de dados de entrada e saída rotulada, e a principal diferença entre classificação e regressão está no tipo de variável alvo que o modelo busca prever. A classificação é utilizada quando a variável de saída é categórica e assume valores discretos que representam classes. Esses algoritmos trabalham com probabilidades de ocorrência de um evento, e a atribuição de uma classe específica pode ser dada pela maior probabilidade aferida, ou por limiares específicos. A regressão é empregada quando a variável de saída é contínua, e o resultado final será um valor em um intervalo numérico (Goodfellow *et al.*, 2016).

b) Divisão das amostras

Para realizar o treinamento de um modelo, amostras suficientes devem ser coletadas e o conjunto de dados deve ser preparado antes do início do processo. Essa preparação envolve dividir os dados em três subconjuntos distintos: treino, validação e teste. O conjunto de treinamento atua no modelo ajustando os parâmetros internos. O modelo passa repetidamente por esses dados durante o treinamento em várias iterações, tentando minimizar uma função de erro ou maximizar acertos.

A base de validação é utilizada para avaliar o desempenho do modelo durante ou após cada iteração de treinamento, sem que haja um aprendizado direto com esses dados, mas com a intenção de estimar o erro de generalização, permitindo que os hiperparâmetros sejam atualizados adequadamente. Por fim, o conjunto de teste é usado para estimar o verdadeiro desempenho de campo de um modelo após a conclusão do processo de aprendizagem, avaliando o modelo de forma final e imparcial (Ripley, 2007).

A taxa de divisão do conjunto de dados é determinada pela quantidade de amostras e da complexidade do problema. Se uma base possui muitos dados, por exemplo, em ordem de centenas de milhares de amostras, pode-se reservar uma proporção menor para teste, pois ainda terá amostras suficientes para uma boa avaliação. Caso contrário, é necessário encontrar um equilíbrio melhor para que todos os conjuntos tenham representatividade. Não existe uma porcentagem de divisão apropriada, mas, normalmente, utiliza-se cerca de 70% dos dados para o treinamento, 15% para a validação, e 15% para o teste.

c) Neurônios e camadas

A definição da arquitetura de uma rede neural, em especial o número de camadas e de neurônios em cada camada, é um fator determinante para o desempenho do modelo. O número de neurônios na camada de entrada é definido pelo número de variáveis ou atributos

de entrada do problema. Para imagens, esse valor deve ser o resultado da multiplicação entre a altura e a largura das imagens para escala de cinza, e novamente multiplicado por três para imagens coloridas de três canais, como é o caso do padrão RGB. Em problemas tabulares, cada coluna da base de dados corresponde a um neurônio de entrada.

Para a camada de saída, o número irá depender do tipo de tarefa a ser realizada, sendo comum a utilização de um único neurônio para regressão ou para classificação binária, e de múltiplos neurônios para classificação multiclasse, onde cada neurônio de saída será a probabilidade de uma classe. O número de unidades ocultas é um parâmetro livre que pode ser ajustado para fornecer o melhor desempenho preditivo (Bishop; Nasrabadi, 2006). Obrigatoriamente, deve estar entre o número de entradas e o número de saídas, e deve ser grande o suficiente para capturar os padrões, mas não tão grande a ponto de superajustar. Em problemas de imagens, o número de neurônios é definido indiretamente pelo tamanho dos *feature maps* e pooling. Uma prática recomendada para definir esses valores seria por testes empíricos de otimização de hiperparâmetros.

Quando se trata de camadas ocultas, as redes rasas com 1 a 2 camadas ocultas podem funcionar bem para problemas menos complexos e dados com menos dimensões ou características. Redes mais profundas de 3 a 5 camadas tendem a oferecer melhor capacidade de generalização e eficiência na representação de funções complexas, e são mais indicadas para tarefas de visão computacional e processamento de linguagem natural.

d) Função de ativação

As funções de ativação são usadas para transformar um sinal de entrada em um sinal de saída, que é então enviado como entrada para a próxima camada. A escolha da função de ativação depende da tarefa específica e da arquitetura da rede, e o processo de aprendizagem pode ser acelerado selecionando a função de ativação apropriada (Rasamoelina *et al.*, 2020). Existem diferentes tipos de funções de ativação. Atualmente, as mais utilizados nas CNNs são:

Função Sigmoide: sua curva parece uma forma de S. A função varia entre 0 e 1, portanto é usada para prever uma probabilidade como saída.

$$f(x) = \frac{1}{1 + e^{-x}} \tag{3.10}$$

Função Tangente Hiperbólica (Tanh): a função tanh tem forma semelhante à função sigmoide, mas o intervalo está entre -1 e 1. A vantagem é que os valores zero serão mapeados próximos de zero e os valores negativos serão mapeados fortemente negativos. Sua definição matemática é:

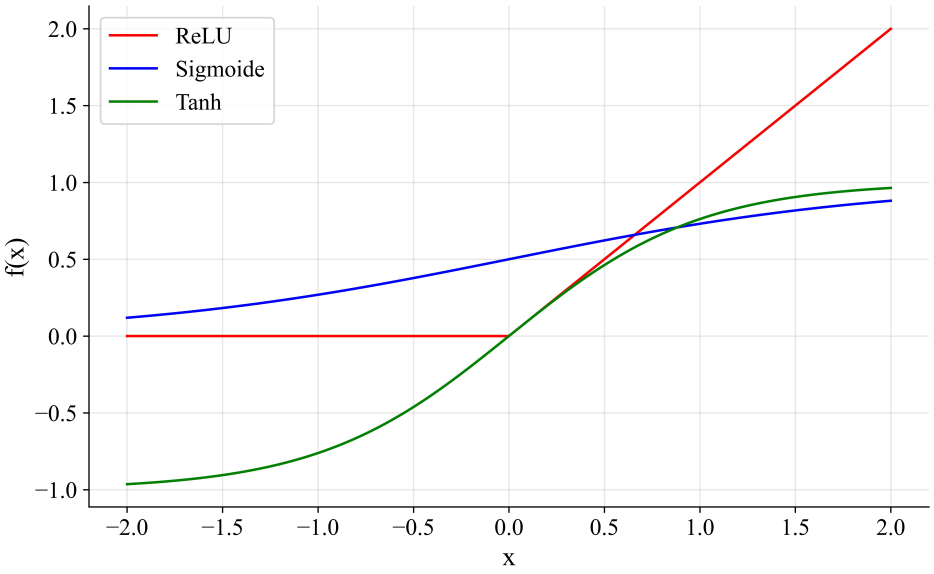
$$f(x) = \tanh(x) = \frac{2}{1 + e^{-2x}} - 1 \tag{3.11}$$

Função Unidade Linear Retificada (ReLU): ReLU é a função de ativação mais utilizada para camadas de convolução. É uma função meio-retificada. É matematicamente definida como:

$$f(x) = max(0, x) = \begin{cases} 0 & x < 0 \\ x & x \geq 0 \end{cases} \tag{3.12}$$

Entender os pontos fortes e fracos de cada função é crucial para selecionar a mais apropriada para uma determinada aplicação de rede neural. Funções Sigmoide e Tanh são adequados para tarefas específicas de classificação e probabilidades, enquanto ReLU é comumente usado em redes profundas modernas. A Figura 15 demonstra os gráficos das funções de ativação.

Figura 15 – Gráfico representativo das funções de ativação



Fonte: Elaborado pela autora.

e) *Batch* e Épocas

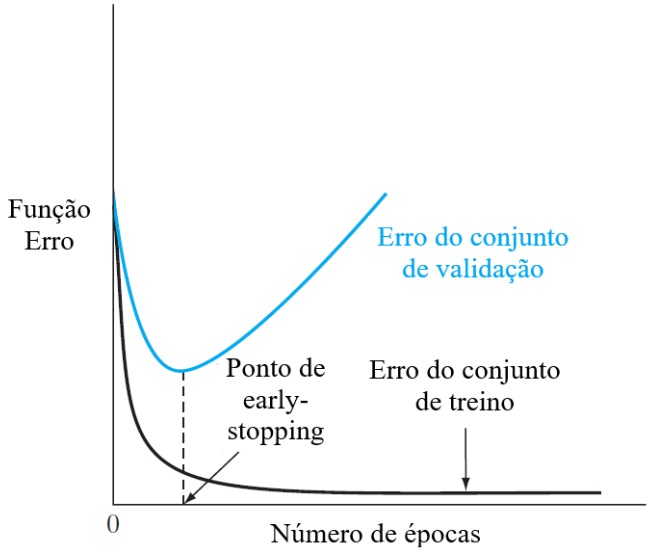
Batch size é o número de dados processados simultaneamente antes da atualização dos pesos da rede durante a estimativa de gradiente. Ao invés dos modelos atualizarem os parâmetros da rede a cada exemplo individual ou somente após todo o conjunto de dados, o treinamento em *mini-batches* faz uma atualização intermediária, conciliando estabilidade no gradiente e eficiência computacional. Valores menores de *batch size* tendem a introduzir

mais ruído, podendo favorecer a capacidade de generalização do modelo, mas com custo computacional maior, enquanto valores maiores fornecem estimativas mais estáveis, mas podem levar a menor poder de generalização e exigir mais memória (Kandel; Castelli, 2020).

Épocas é um hiperparâmetro que determina o número de passagens completas para uma rede neural durante o treinamento. Durante cada época, o modelo processa todos os exemplos disponíveis, geralmente em *mini-batches*, e ajusta seus pesos. Tradicionalmente, o número de épocas é grande, permitindo que o procedimento de aprendizado continue até que o erro do modelo seja adequadamente minimizado (Zhang *et al.*, 2016). Poucas épocas podem resultar em casos onde a rede não aprende suficientemente os padrões dos dados, porém épocas demais também podem acarretar na memorização dos dados de treino, perdendo capacidade de generalização.

Não existe uma pré-definição do número correto de épocas para todos os conjuntos de dados, mas é possível determinar uma aproximação ideal utilizando a curva de aprendizado, um gráfico de linhas que mostra a evolução dos erros nas bases de treino e validação com o acréscimo de épocas. Para muitos dos algoritmos o erro de treinamento é sempre uma função decrescente, no entanto, o erro calculado em relação ao conjunto de validação frequentemente apresenta uma diminuição inicial, seguida por um aumento à medida que a rede começa a sofrer *over-fitting*. O ponto de interrupção é obtido utilizando a estratégia de *early-stopping*, onde o treinamento é encerrado após o erro nos dados de validação atingir a saturação, que pode ser definida em número de épocas subsequentes com nenhuma melhora ou piora nos erros. A Figura 16 ilustra a aplicação do método.

Figura 16 – Exemplificação da regra de *early-stopping*



Fonte: Haykin (2009). Adaptado pela autora.

f) Taxa de aprendizagem

A taxa de aprendizagem determina o tamanho do passo que o modelo dá na direção oposta ao gradiente do erro em cada iteração, controlando assim a velocidade com que os pesos da rede são atualizados. É muitas vezes descrita nas equações da literatura como η, e seus valores são pequenos e positivos, geralmente variando entre 0.0 e 1.0. Se for muito grande ($\eta > 0.1$), a curva de aprendizado apresentará oscilações violentas, pois os pesos saltam demais no espaço da função de perda. Se a taxa de aprendizagem for muito baixa ($\eta < 10^{-4}$), o aprendizado prossegue lentamente e o aprendizado pode ficar estagnado em mínimos locais (Bengio, 2012).

A taxa de aprendizagem pode ser escolhida por experimentação empírica, mas geralmente é melhor escolhê-la monitorando as curvas de aprendizagem que correlacionam a função objetivo com as iterações do modelo. É um dos hiperparâmetros mais sensíveis no treinamento de redes neurais, exigindo cuidadosa escolha e, muitas vezes, uso de técnicas adaptativas e de otimização.

3.4 Métricas de Performance

A avaliação de modelos de aprendizado de máquina requer o uso de métricas de desempenho adequadas, capazes de refletir de maneira confiável a qualidade das previsões. A escolha das métricas depende diretamente da natureza do problema, podendo ser de classificação ou de regressão. Além de quantificar a qualidade do modelo, as métricas de performance desempenham papel crucial na comparação entre diferentes abordagens e na orientação do ajuste de hiperparâmetros. A definição correta das avaliações é uma etapa indispensável para a validação de modelos, garantindo a relevância prática das soluções propostas. Normalmente, o interesse principal está na performance de um algoritmo em dados que ele não viu antes, pois isso determina quão bem ele funcionará quando implantado no mundo real. Portanto, a avaliação final de um modelo será usando sempre um conjunto de dados de teste.

3.4.1 Classificação

Em problemas de classificação, a performance está relacionada à capacidade do modelo de prever corretamente rótulos aos dados de entrada. As métricas usados para avaliar a robustez de um classificador são acurácia, precisão e recall. **Acurácia** diz a proporção de previsões corretas (positivas e negativas) em relação ao total de amostras, medindo o quão frequentemente o modelo acerta. **Precisão** calcula a proporção de instâncias classificadas como positivas que realmente são positivas, estimando a confiança nas previsões positivas. Por fim, o **Recall** calcula a proporção de instâncias realmente positivas que o modelo conseguiu identificar, ou seja, a capacidade do modelo de encontrar

todos os positivos. Frequentemente, Precisão e Recall são analisados juntos. As fórmulas para calcular esses valores são apresentadas abaixo:

$$Acurácia = \frac{VP + VN}{VP + VN + FP + FN} \tag{3.13}$$

$$Precisão = \frac{VP}{VP + FP} \tag{3.14}$$

$$Recall = \frac{VP}{VP + FN} \tag{3.15}$$

onde:

VP - Verdadeiro Positivo.
VN - Verdadeiro Negativo.
FP - Falso Positivo.
FN - Falso Negativo.

3.4.2 Regressão

Em tarefas de regressão, nas quais o objetivo é prever valores contínuos, a maneira de avaliar o modelo é medir a correlação entre os valores reais e os previstos. A avaliação geralmente se baseia em medidas de erro, tais como o MSE, o RMSE, MAE e MAPE, que capturam diferentes aspectos da precisão e robustez das previsões. O *Mean Squared Error* (Erro Quadrático Médio) (MSE) calcula a média dos quadrados dos erros de previsão, penalizando mais fortemente erros grandes devido à elevação ao quadrado. O *Root Mean Squared Error* (Raiz do Erro Quadrático Médio) (RMSE) é a raiz quadrada do MSE, mas retorna os valores na mesma unidade da variável prevista, facilitando a interpretação.

O *Mean Absolute Error* (Erro Absoluto Médio) (MAE) é a média dos valores absolutos dos erros, e representa o erro médio em termos absolutos, sem exagerar o peso de grandes desvios. Por fim, o *Mean Absolute Percentage Error* (Erro Percentual Absoluto Médio) (MAPE) mostra o erro médio do modelo em termos percentuais. As definições de cada métrica se encontram nas equações abaixo:

$$MSE = \frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{y}_i)^2 \tag{3.16}$$

$$RMSE = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{y}_i)^2} = \sqrt{MSE} \tag{3.17}$$

$$MAE = \frac{1}{n}\sum_{i=1}^{n}|y_i - \hat{y}_i| \tag{3.18}$$

$$MAPE = \frac{100}{n} \sum_{i=1}^{n} \frac{y_i - \hat{y}_i}{y_i} \tag{3.19}$$

onde:

- y_i - Valor verdadeiro.
- $\hat{y}_i$ - Valor previsto.
- n - Número total de observações.

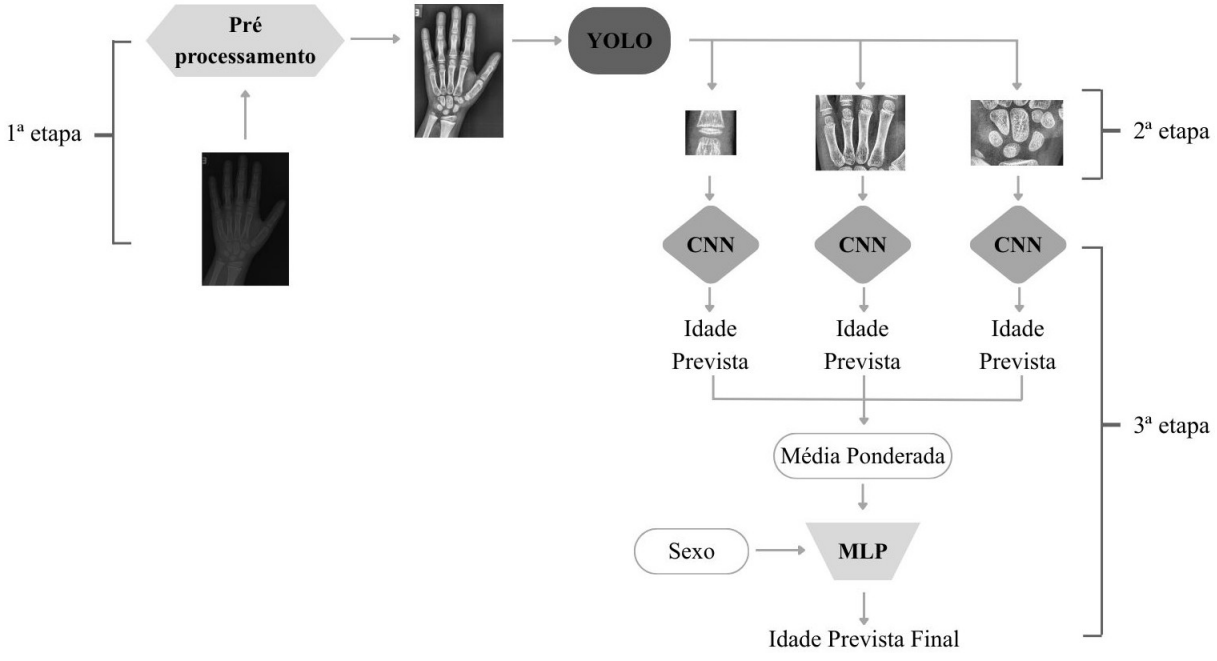
4 MATERIAIS E MÉTODOS

4.1 Introdução

De modo geral, modelos de aprendizado de máquina buscam estabelecer um mapeamento entre um conjunto de entradas e a forma desejada de saída. No contexto da estimativa da idade óssea a partir de radiografias da mão, a imagem constitui a entrada do sistema, enquanto a idade óssea estimada representa a saída. Como a variável a ser predita é contínua, a formulação do problema como uma tarefa de regressão apresenta-se como a abordagem mais adequada para modelar a relação entrada–saída.

Este capítulo descreve a metodologia adotada para a predição da idade óssea a partir das imagens radiográficas, detalhando os recursos empregados, as etapas de pré-processamento e a arquitetura das redes neurais artificiais utilizadas. A fim de sintetizar as etapas e componentes desenvolvidos ao longo deste trabalho, apresenta-se na Figura 17 uma visão geral do fluxo completo do método proposto.

Figura 17 – Fluxo do método completo



Fonte: Elaborado pela autora.

4.2 Banco de dados

A idade óssea constitui uma medida utilizada para quantificar o estágio de desenvolvimento esquelético em crianças e adolescentes. Essa métrica difere da idade cronológica, uma vez que é expressa em uma escala que varia de 0 a 228 meses. Em condições de crescimento consideradas normais, crianças do mesmo sexo e com idades cronológicas idênticas tendem a apresentar idades ósseas semelhantes (Cunningham *et al.*, 2016).

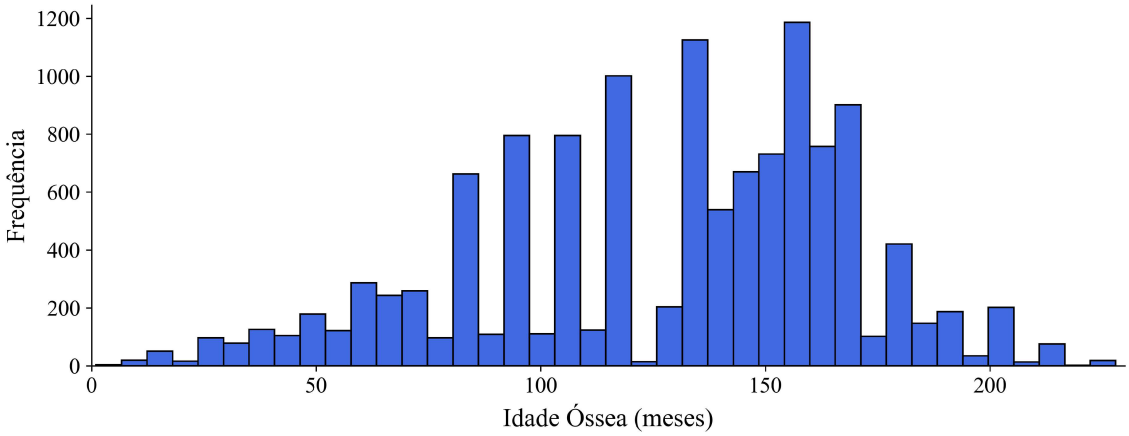
O desenvolvimento de modelos de aprendizado profundo depende fortemente da disponibilidade e da qualidade dos dados utilizados no processo de treinamento e validação. Nesse contexto, a construção e a caracterização do banco de dados constituem etapas fundamentais, pois a representatividade e a diversidade das amostras impactam diretamente no desempenho e na capacidade de generalização do modelo. Para o treinamento do modelo proposto neste trabalho, foram utilizadas duas bases de dados públicas distintas: a RSNA e o RHPE.

O conjunto de dados da *Radiological Society of North America* (Sociedade de Radiologia da América do Nort) (RSNA) foi desenvolvido com o objetivo de apoiar o avanço e a avaliação de técnicas automatizadas de estimativa da idade óssea. Ele é composto por 12.611 radiografias da mão, obtidas majoritariamente em dois hospitais pediátricos nos Estados Unidos: o Children's Hospital Colorado e o Lucile Packard Children's Hospital (RSNA, 2017).

Além das radiografias, é disponibilizado um arquivo em formato CSV contendo a idade óssea aferida por radiologistas, bem como a informação de gênero dos indivíduos. A anotação das idades foi realizada por seis especialistas médicos, com o valor final atribuído a cada exame calculado a partir da média ponderada das avaliações.

A base contempla imagens de participantes de ambos os sexos, sendo 5.778 (46%) correspondentes a participantes do sexo feminino e 6.833 (54%) do sexo masculino. As imagens foram adquiridas por diferentes equipamentos de raio-X, apresentando variação de resolução que vai de 800 $\times$ 1.011 pixels até 2.460 $\times$ 2.970 pixels. A Figura 18 apresenta a distribuição de frequência da idade óssea no conjunto de dados.

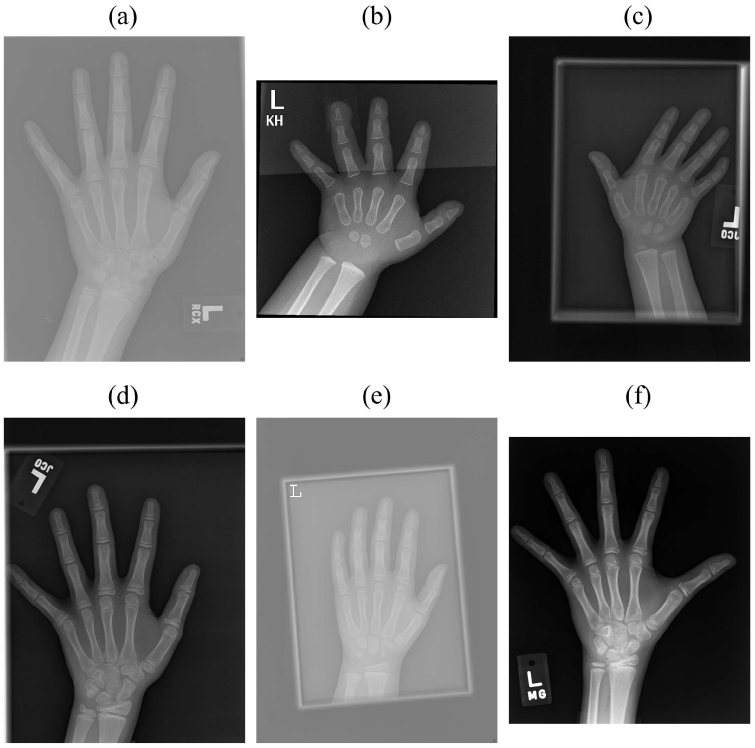
Figura 18 – Distribuição do dataset RSNA



Fonte: Elaborado pela autora.

A Figura 19 apresenta alguns exemplos de radiografias da mão provenientes da base RSNA, ilustrando a diversidade de qualidade, posição e características anatômicas presentes no conjunto de dados utilizado como referência neste trabalho.

Figura 19 – Exemplos de imagens presentes na base RSNA

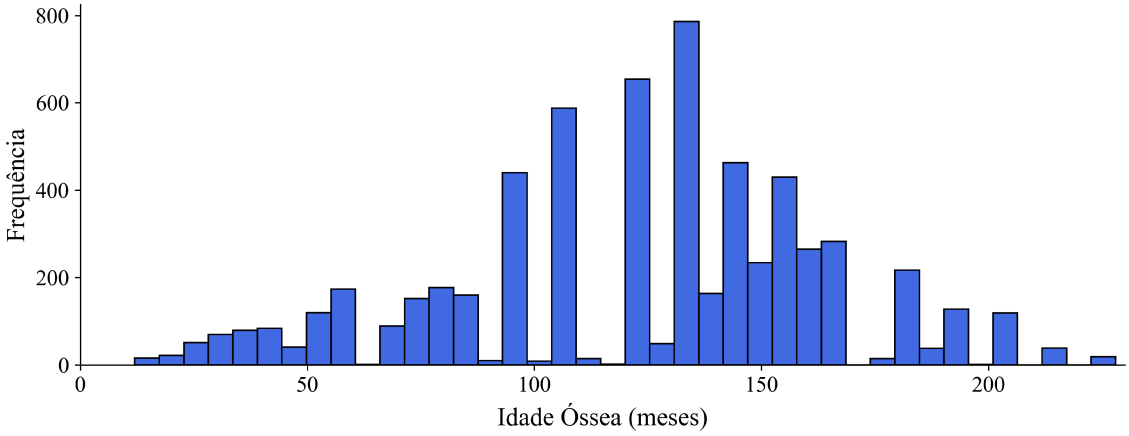


Fonte: Elaborado pela autora.

O estudo de (Escobar *et al.*, 2019) apresentou um novo conjunto de dados de referência, denominado *Radiological Hand Pose Estimation* (Estimativa Radiológica da Postura da Mão) (RHPE), desenvolvido a partir da análise local de regiões de interesse em radiografias da mão. Essa base é composta por 6.204 radiografias, provenientes de uma população com características distintas das bases até então disponíveis para a tarefa de avaliação da idade óssea, o que garante maior variabilidade e contribui para uma melhor capacidade de generalização dos modelos.

O banco inclui imagens de mãos direitas e esquerdas, de pacientes do sexo masculino e feminino, com idades compreendidas entre 0 e 228 meses, e as anotações de idade óssea foram realizadas por dois radiologistas especialistas. Em termos de distribuição por sexo, 54% (3.526) das radiografias correspondem a pacientes do sexo feminino, enquanto 46% (2.678) pertencem a pacientes do sexo masculino. Adicionalmente, a base disponibiliza também um arquivo CSV informações referentes à idade óssea e ao sexo dos pacientes. As imagens apresentam variação nas dimensões, indo da menor resolução de 973 × 1.192 pixels até a maior de 8.480 × 6.960 pixels. A Figura 20 apresenta a distribuição nesse conjunto de dados.

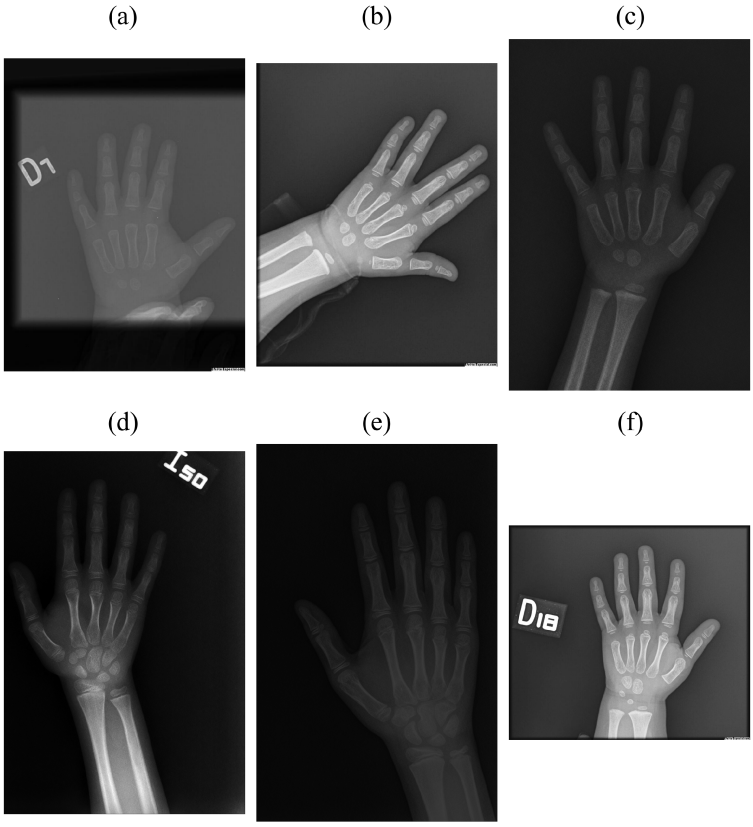
Figura 20 – Distribuição do dataset RHPE



Fonte: Elaborado pela autora.

A Figura 21 exibe amostras de radiografias da base RHPE, evidenciando a variabilidade típica do banco institucional e destacando diferenças de aquisição e padrões de posicionamento em relação à base RSNA.

Figura 21 – Exemplos de imagens presentes na base RHPE



Fonte: Elaborado pela autora.

Uma análise comparativa dos gráficos para os conjuntos de dados RSNA e RHPE mostra que ambos possuem um formato de distribuição muito semelhante, onde a concentração de dados não é uniforme. Há poucas amostras nas idades mais jovens (abaixo de 50 meses) e uma quantidade significativamente maior no período da adolescência, entre 120 e 180 meses (10 a 15 anos), sendo um reflexo direto da prática médica do mundo real. A principal razão pela qual se avalia a idade óssea é para investigar possíveis distúrbios de crescimento, e a maioria dessas preocupações se tornam mais evidente durante a puberdade e pré-puberdade. Ainda assim, se complementam resultando em uma base de dados mais diversificada, abrangendo variações de diferentes populações e equipamentos. Essa combinação é estratégica para construir modelos de inteligência artificial mais robustos e capazes de funcionar bem em novos cenários clínicos.

4.3 Softwares e Bibliotecas

4.3.1 Ambiente de desenvolvimento

O desenvolvimento do presente trabalho foi realizado utilizando a linguagem de programação Python, em conjunto com o ambiente de desenvolvimento integrado Visual Studio Code (VS Code). A escolha do Python se deve à sua ampla adoção na comunidade científica e de aprendizado de máquina, bem como à disponibilidade de bibliotecas especi-

alizadas para processamento de imagens, manipulação de dados e construção de modelos de aprendizado profundo. O VS Code, por sua vez, oferece recursos avançados de edição, depuração e integração com sistemas de controle de versão, proporcionando um fluxo de trabalho eficiente e organizado. Essa combinação possibilitou a implementação de rotinas complexas de pré-processamento, treinamento e avaliação de modelos de forma reprodutível e modular, favorecendo tanto a produtividade quanto a manutenção do código.

4.3.2 Bibliotecas e pacotes

Nesta subseção são apresentadas as principais bibliotecas e pacotes Python empregados no desenvolvimento do projeto. Tais ferramentas foram utilizadas para etapas como manipulação de dados, pré-processamento de imagens, construção e treinamento de modelos, visualização gráfica e avaliação de desempenho. A lista a seguir descreve individualmente cada biblioteca empregada e suas funções.

a) Numpy

Biblioteca fundamental para computação científica, oferecendo estruturas de dados eficientes, como arrays multidimensionais, além de funções matemáticas e operações vetorizadas. Sua utilização permite manipulações rápidas e eficientes de grandes volumes de dados numéricos, sendo amplamente empregada em pré-processamento de imagens e operações matriciais em aprendizado de máquina.

b) Matplotlib

Uma biblioteca de visualização que possibilita a criação de gráficos estáticos, animados e interativos. Ela é utilizada para a análise exploratória de dados, acompanhamento do treinamento de modelos e apresentação de resultados de forma visualmente clara e personalizável, contribuindo para a interpretação e comunicação dos achados.

c) OpenCV

O OpenCV *Open Source Computer Vision Library* (Biblioteca de Visão Computacional de Código Aberto) é uma biblioteca voltada para processamento e análise de imagens e vídeos. Ela oferece funções para leitura, transformação, segmentação e manipulação de imagens, sendo essencial para tarefas de pré-processamento.

d) Pandas

Biblioteca especializada em manipulação e análise de dados tabulares estruturados. Por meio de suas estruturas de dados, é possível organizar, filtrar, agregar e transformar grandes conjuntos de informações de maneira eficiente, permitindo importar e exportar dados de vários formatos de arquivos de tabelas e consultas à banco de dados.

e) Pytorch

O PyTorch é uma biblioteca de aprendizado profundo que fornece uma interface

flexível para construção, treinamento e avaliação de redes neurais. Sua arquitetura dinâmica de grafos computacionais facilita a implementação de modelos complexos e o acompanhamento de gradientes.

f) Scikit-learn

Biblioteca de aprendizado de máquina que oferece algoritmos de classificação, regressão, agrupamento e pré-processamento de dados, construído sobre NumPy, Pandas e Matplotlib. Sua facilidade de integração e conjunto abrangente de métricas permitem validação de modelos e análise estatística de dados.

g) Seaborn

O Seaborn é uma biblioteca de visualização estatística baseada no Matplotlib. Ela facilita a criação de gráficos complexos e esteticamente consistentes, como mapas de calor, distribuições e boxplots, sendo útil na análise exploratória de dados.

h) Ultralytics

O Ultralytics fornece implementações avançadas da família de modelos YOLO, voltadas para detecção de objetos em imagens, com compatibilidade com frameworks de aprendizado profundo como PyTorch. Sua utilização permite treinar e aplicar redes de alta performance para identificar estruturas ou objetos específicos, oferecendo suporte a diferentes resoluções de entrada e facilidades para integração com pipelines de visão computacional.

i) Optuna

Biblioteca de otimização de hiperparâmetros que automatiza a busca por combinações ideais de parâmetros em modelos de aprendizado de máquina. Ele utiliza técnicas de otimização inteligente para aumentar a eficiência e desempenho de modelos computacionais. O espaço de busca de hiperparâmetros é definido dinamicamente durante a execução do código, permitindo flexibilidade na experimentação. A biblioteca utiliza algoritmos de otimização sequencial para sugerir novas combinações de parâmetros com base nos resultados de avaliações anteriores, equilibrando exploração e exploração do espaço de busca. Além disso, o Optuna oferece recursos para paralelização de experimentos, registro automático de resultados e integração com frameworks populares como PyTorch e TensorFlow.

j) Albumentations

Ferramenta amplamente utilizada em visão computacional por sua eficiência e flexibilidade no tratamento conjunto de imagens para tarefas de detecção de objetos. Diferentemente de bibliotecas tradicionais de aumento de dados, a Albumentations garante que, ao aplicar uma transformação geométrica na imagem (como rotação, translação ou escala), as coordenadas das caixas delimitadoras sejam ajustadas automaticamente e de forma precisa, preservando a correspondência entre imagem e rótulo.

4.3.3 Ferramentas auxiliares

Além das bibliotecas de software utilizadas diretamente na implementação, este trabalho também fez uso de ferramentas auxiliares essenciais para o preparo dos dados e para o processamento eficiente dos modelos. Nesta subseção são descritas a ferramenta de anotação de imagens e o recurso de processamento acelerado utilizados ao longo do projeto.

a) CVAT

O CVAT *Computer Vision Annotation Tool* (Ferramenta de Anotação de Visão Computacional) é uma ferramenta de código aberto para anotação de imagens e vídeos, permitindo a marcação de objetos, regiões de interesse e pontos de referência. Ele é utilizado para criar dados anotados de referência em tarefas de visão computacional, oferencendo uma variedade de opções de rotulagem.

b) GPU

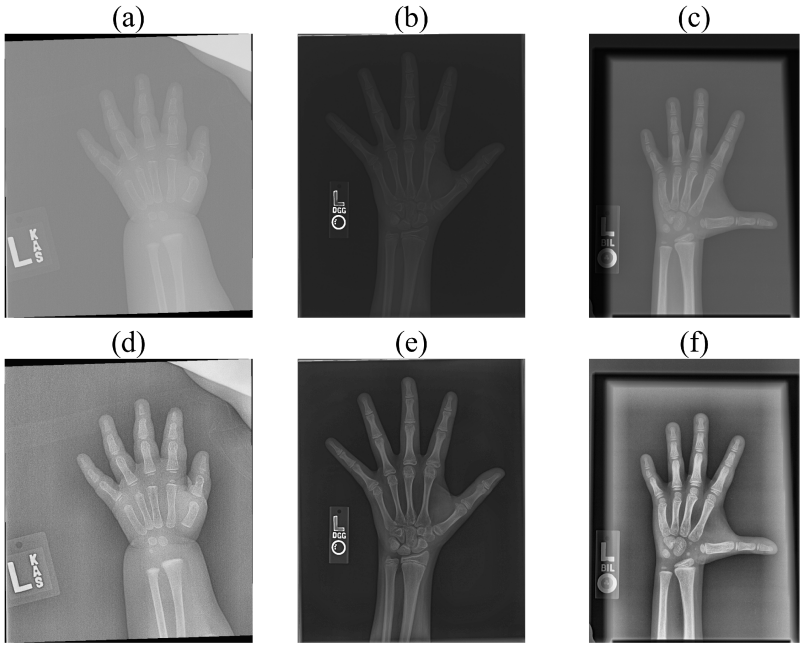
A GPU (Graphics Processing Unit) é um componente de hardware especializado em processamento paralelo, originalmente desenvolvidos para aplicações gráficas. Seu uso permite acelerar significativamente operações computacionais complexas, como o treinamento de redes neurais e o processamento intensivo de imagens. Todos os experimentos foram implementados em uma GPU com especificações NVIDIA GeForce GTX 1060 6GB.

4.4 Pré Processamento

Diversos fatores influenciam a aparência final de imagens radiográficas. Os valores do pixels são afetados por parâmetros técnicos utilizados durante a aquisição da radiografia, pela orientação da mão em relação ao detector de imagem e pelo nível geral de luminosidade da região registrada. Nos conjuntos de dados observa-se uma ampla variação na intensidade das imagens decorrente da natureza multi-institucional do material, coletado com equipamentos e protocolos distintos. Assim, torna-se necessário reduzir as variações de intensidade que não estão relacionadas às características estruturais de interesse, de modo a evitar que ruídos de aquisição interfiram na análise computacional.

Um método amplamente reconhecido para esse fim é a equalização de histogramas, utilizada para uniformizar distribuições de intensidade em conjuntos de imagens. Neste trabalho, foi empregada a técnica de CLAHE, a fim de normalizar as diferentes variações de tons de cinza e realçar detalhes relevantes para a tarefa proposta. A Figura 22 apresenta algumas imagens após a aplicação desse procedimento na base de dados.

Figura 22 – Exemplares de radiografias disponíveis na base. a), b) e c) Imagens originais. d), e) e f) Correção com CLAHE



Fonte: Elaborado pela autora.

4.5 Extração das regiões de interesse

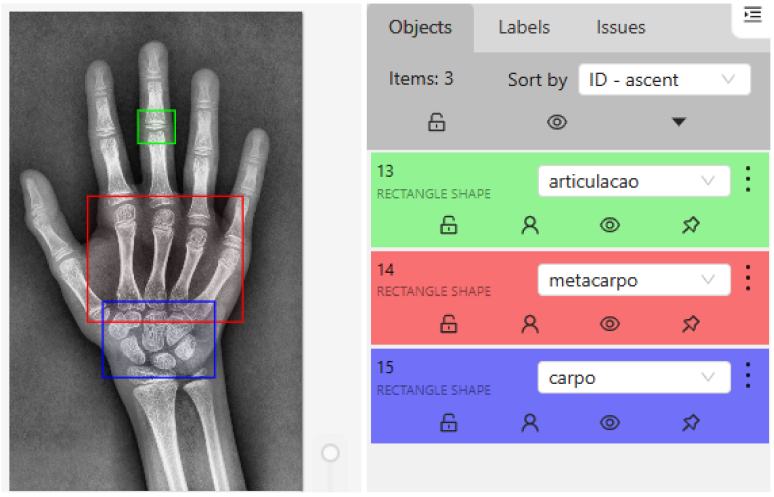
Selecionar de forma adequada a região da imagem utilizada para inferência é fundamental, pois contribui para eliminar variações visuais que não têm relação com a idade óssea. Nas radiografias, o fundo da imagem não fornece informações úteis para essa estimativa e, por isso, pode ser descartado sem comprometer o desempenho do modelo. Além disso, o padrão de ossificação tende a ser semelhante em diferentes partes da mão, o que torna algumas informações visuais repetitivas e, portanto, desnecessárias para a análise. Assim, ao definir as regiões de interesse, deve-se considerar o equilíbrio entre eliminar redundâncias, preservar as características únicas e manter apenas os aspectos realmente relevantes para a predição da idade óssea.

Para viabilizar a extração automática das regiões de interesse, foi empregado um modelo da família YOLO, especificamente a versão 10 disponibilizada pela Ultralytics. O primeiro passo para treinar o modelo é a anotação manual das regiões relevantes da mão, que servirá como base de aprendizado supervisionado. Esse processo foi realizado com auxílio da plataforma CVAT (Computer Vision Annotation Tool).

Para este trabalho, foram rotuladas 204 imagens, selecionadas manualmente de forma criteriosa a fim de representar a diversidade existente no banco de dados. Nessa seleção, buscou-se contemplar diferentes condições, como variações no posicionamento e orientação da mão, níveis distintos de brilho e contraste, diferenças na proporção entre a

área da mão e o tamanho total da imagem, além da presença de bordas e identificadores dos exames. No CVAT, cada radiografia é então carregada e a área de interesse é delimitada manualmente por meio de caixas delimitadoras. Essas anotações são associadas a rótulos específicos, permitindo ao modelo aprender a identificar automaticamente essas regiões em novas imagens. O processo de rotulação manual e geração de caixas delimitadoras pode ser visto na Figura 23.

Figura 23 – Interface do CVAT para rotulação de imagens



Fonte: Elaborado pela autora.

As regiões escolhidas para o desenvolvimento do trabalho foram três: **(i)** a articulação entre a falange média e proximal, **(ii)** os quatro ossos do metacarpo localizados na palma da mão e **(iii)** os ossos do carpo. Após a conclusão da etapa de anotação, os arquivos com as coordenadas dessas regiões foram exportados em formato compatível com o YOLOv10, constituindo, junto com as imagens originais, o conjunto de dados de treinamento. Cada imagem recebe um arquivo de rótulo correspondente no formato .txt, de modo que a estrutura dos diretórios seguiu o padrão: */data/train/images/00001.png* para a imagem e */data/train/labels/00001.txt* para o respectivo rótulo.

O arquivo de rótulo contém uma linha para cada região anotada dentro da imagem, composta por cinco valores separados por espaços, representando as informações necessárias para que o modelo localize e classifique os objetos de forma precisa. O valor de *classe* indica a região anotada, *x_centro* e *y_centro* representam as coordenadas do centro da caixa delimitadora, e *largura* e *altura* correspondem às dimensões da caixa delimitadora, sendo todos esses valores normalizados em relação à largura e altura total da imagem. A normalização é realizada dividindo a coordenada original pelo tamanho da imagem correspondente, resultando em valores entre 0 e 1, garantindo que as anotações permaneçam consistentes mesmo quando as imagens são redimensionadas. A Tabela 2 exemplifica o formato do arquivo de anotação:

Tabela 2 – Exemplo da formatação de saída das anotações.

<classe>	<x_centro>	<y_centro>	<largura>	<altura>
0	0.510735	0.229892	0.129617	0.086406
1	0.470988	0.507748	0.407023	0.212016
2	0.408085	0.622367	0.213246	0.116426

O conjunto total de imagens anotadas foi dividido em duas partições: 70% destinadas ao treinamento e 30% reservadas para validação, assegurando que o modelo fosse avaliado em dados não vistos durante o aprendizado, sendo fundamental para mensurar a capacidade de generalização. Para ampliar a variabilidade do conjunto de treinamento e tornar o modelo mais robusto a diferentes condições de captura, aplicou-se uma etapa de aumento de dados, gerando duas novas versões de cada imagem original combinando diversas variações aplicadas de forma aleatória, com auxílio da biblioteca Albumentations. As transformações disponibilizadas foram: **(i)** Espelhamento horizontal completo, permitindo que o modelo reconheça a mão independentemente da lateralidade; **(ii)** Rotação aleatória no intervalo de -10° a 10°, simulando pequenas variações de posicionamento durante o exame radiográfico; **(iii)** Variação aleatória de escala de até 10%, ajustando a proporção entre a mão e a área total da imagem; **(iv)** Modificação controlada nos parâmetros de brilho (±20%), contraste (±10%) e saturação (±20%), de modo a representar variações de iluminação e qualidade de aquisição da imagem.

Após essa etapa, iniciou-se o processo de configuração e treinamento do modelo. O YOLOv10 fornece cinco versões escalonadas: YOLOv10n (nano), YOLOv10s (pequeno), YOLOv10m (médio), YOLOv10l (grande) e YOLOv10x (extragrande). Cada variação do YOLOv10 usa a mesma arquitetura base, mas difere em profundidade da rede, número de canais por camada, número total de parâmetros, e uso de memória e tempo de inferência. Optou-se pela versão pequena (s) do modelo, que oferece um bom equilíbrio entre desempenho computacional e precisão na detecção, sendo especialmente adequada para conjuntos de dados de tamanho moderado e aplicações em pesquisa.

O treinamento foi conduzido localmente utilizando GPU, o que possibilitou maior eficiência no processamento de lotes e convergência mais rápida do modelo. As imagens são redimensionadas automaticamente na entrada do modelo, e os principais hiperparâmetros adotados foram os seguintes:

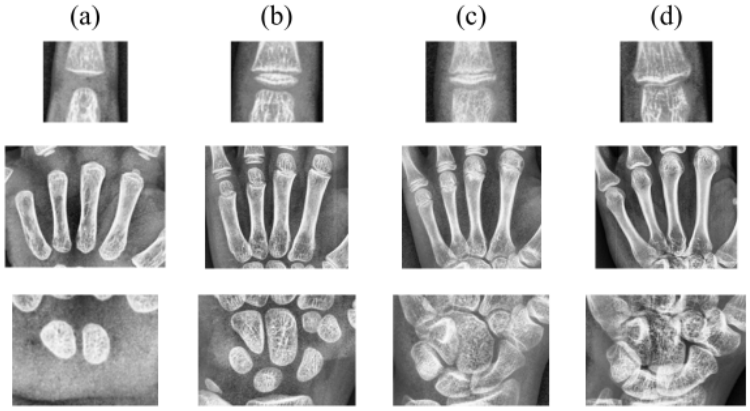
Tamanho da imagem	-	800×800 px
Número de épocas	-	200
Batch size	-	8
Taxa de aprendizado	-	10^{-3}
Early stopping	-	20

4.6 Detecção da Idade Óssea

Com o modelo YOLOv10 devidamente treinado e validado, tornou-se possível aplicar o detector a toda a base de imagens radiográficas, automatizando o processo de localização e recorte das regiões de interesse previamente definidas. Assim, para cada imagem original da base, o modelo realiza a inferência e identifica, por meio das caixas delimitadoras aprendidas durante o treinamento, as áreas correspondentes às três regiões anatômicas definidas.

Com esses resultados, inicia-se a fase de predição da idade óssea, conduzida por uma rede neural. Nessa etapa, as regiões extraídas são utilizadas como entradas do modelo regressivo, que tem como objetivo estimar, de forma contínua, a idade óssea do indivíduo, extraindo características discriminativas relacionadas à morfologia e densidade óssea dos diferentes estágios de ossificação. Para ilustrar visualmente o processo de desenvolvimento esquelético considerado neste estudo, a Figura 24 apresenta um exemplo da evolução da ossificação nas regiões extraídas ao longo de diferentes faixas etárias. Observa-se que, com o avanço da idade, há uma progressiva fusão e aumento na densidade e definição das estruturas ósseas.

Figura 24 – Evolução da ossificação nas regiões extraídas: articulação, metacarpo e carpo, respectivamente. a) Pacientes com 18 meses; b) 84 meses; c) 162 meses e d) 204 meses



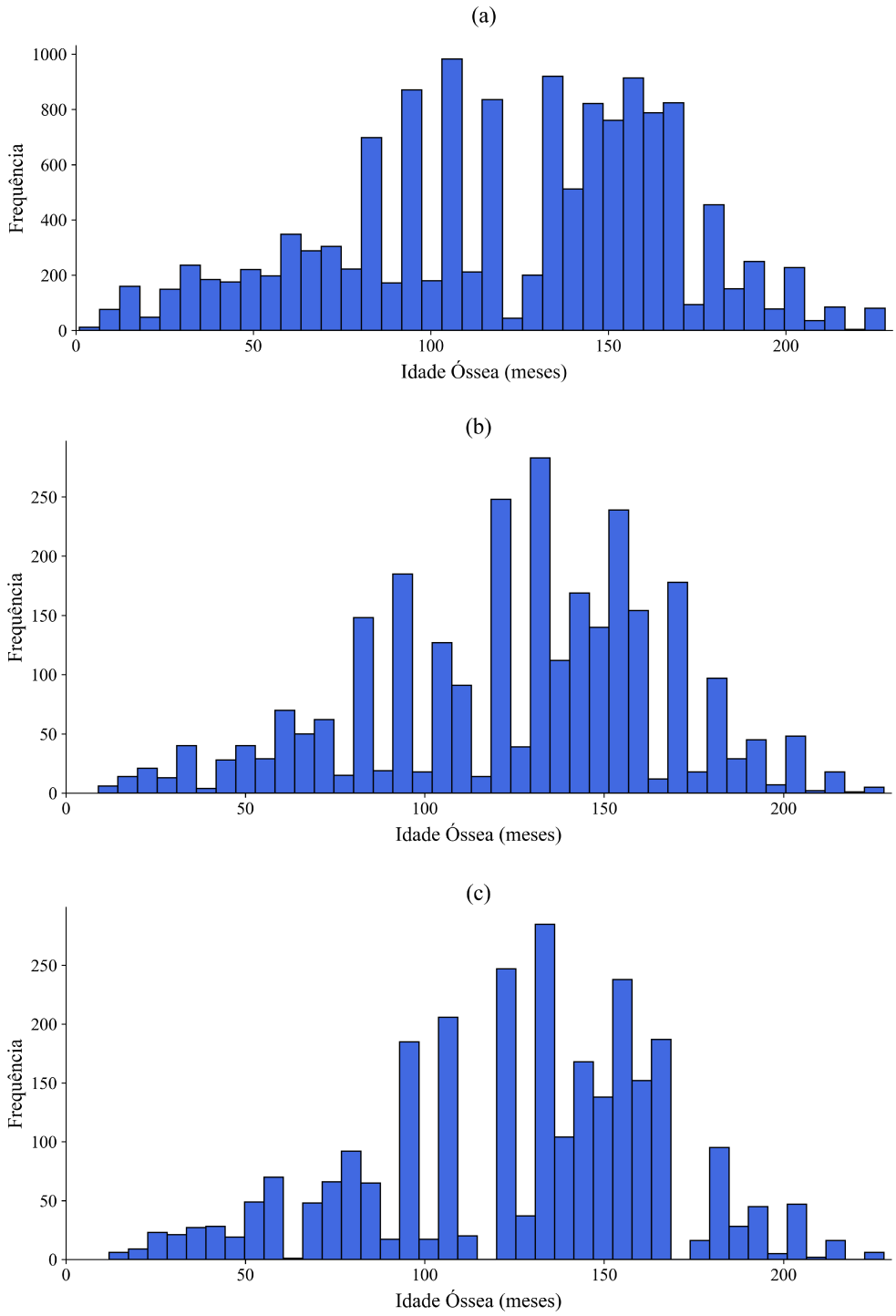
Fonte: Elaborado pela autora.

O particionamento do conjunto de dados utilizado para o treinamento e avaliação do modelo foi realizado de forma estratificada, considerando a variável alvo, de modo a garantir que todas as divisões apresentassem uma distribuição semelhante de idades. Dados de toda a faixa etária foram incluídos para treinamento no estudo. Assim, 70% das amostras foram destinadas ao conjunto de treinamento, enquanto 15% foram reservadas para validação e os 15% restantes para teste.

Para lidar com o desbalanceamento da base de dados, foi empregado *Data Augmentation* no conjunto de treino, aumentando artificialmente as amostras das classes

minoritárias, e subamostragem nas classes majoritárias. Para as classes minoritárias, cada imagem original gerou 3 cópias adicionais com modificações: **(i)** Espelhamento horizontal; **(ii)** Rotações aleatórias de até ±30°; **(iii)** Translações no eixo x de $\pm 15px$ e no eixo y de $\pm 10px$; **(iv)** Ajustes de brilho e contraste de ±20%. As classes majoritárias foram limitadas em 800 amostras. A distribuição dos conjuntos pode ser visto na Figura 25.

Figura 25 – Distribuição das amostras em cada base. a) Treino; b) Validação; c) Teste



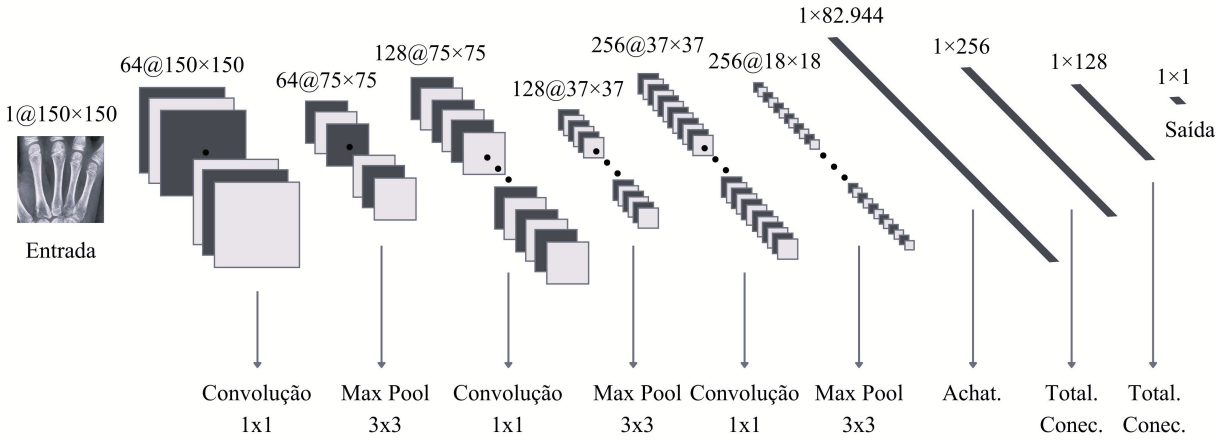
Fonte: Elaborado pela autora.

Ao final desse processo, todas as imagens originais e aumentadas foram consolidadas em um diretório final, resultando em um conjunto com 13.815 amostras para treino, 2.838 para validação, e 2.785 para teste. Todas as partições possuem histogramas comparáveis, garantindo que não haja concentração excessiva de determinadas faixas etárias, permitindo que o modelo aprenda de forma consistente os padrões de ossificação para todas as idades e seja avaliado de maneira justa, refletindo sua capacidade de generalização. As imagens recortadas de todas as regiões foram redimensionadas reduzindo o tamanho para 150 × 150 pixels e convertidas para escala de cinza, assegurando compatibilidade com a rede neural convolucional regressora utilizada na etapa seguinte.

O modelo de predição da idade óssea desenvolvido neste trabalho se baseia em uma arquitetura composta por três redes neurais convolucionais independentes, cada uma especializada em processar uma região específica. Cada rede convolucional segue a mesma arquitetura, consistindo em três blocos convolucionais sequenciais. Cada bloco é composto por uma camada convolucional seguida de função de ativação ReLU e operação de Max Pooling, com o objetivo de extrair características relevantes das imagens e reduzir gradualmente suas dimensões espaciais. O primeiro bloco convolucional recebe a imagem em escala de cinza e aplica 64 filtros de tamanho 3×3, seguidos de ativação ReLU e uma operação de *Max Pooling*, resultando em *feature maps* de dimensão 75×75. O segundo bloco, com 128 filtros de tamanho 3×3, processa os 64 canais provenientes da etapa anterior, reduzindo a dimensão espacial para 37×37. Por fim, o terceiro bloco convolucional utiliza 256 filtros também de 3×3, extraindo representações mais complexas e condensadas da imagem, com *feature maps* finais de tamanho 18×18.

A saída convolucional é então achatada, transformando-se em um vetor unidimensional de tamanho 1×82.944, resultante da multiplicação entre o tamanho dos *feature maps* e a quantidade de filtros da última camada convolucional. Em seguida, são aplicadas transformações lineares, responsáveis por combinar as características extraídas nas etapas anteriores e gerar representações significativas dos dados. Essas camadas totalmente conectadas possuem dimensões de 1×256 e 1×128, respectivamente, seguidas por funções de ativação ReLU. Por fim, uma camada linear de saída única produz a predição da idade óssea correspondente a cada região analisada. A Figura 26 apresenta as camadas da rede desenvolvida.

Figura 26 – Gráfico arquitetônico da CNN



Fonte: Elaborado pela autora.

Após a inferência de cada região, as predições individuais de cada rede são combinadas por meio de uma média ponderada, na qual os pesos são calculados com base no Erro Quadrático (EQ) individual de cada predição em relação à idade real do paciente. Essa estratégia permite que predições mais confiáveis exerçam maior influência na estimativa final antes da etapa de refinamento. Os erros são obtidos pela equação (20):

$$EQ_i = (\hat{y}_i - y)^2 \tag{4.1}$$

Em seguida, cada erro é transformado em um peso (w_i) por meio da função *softmax*, aplicada sobre os valores negativos dos erros, de forma que previsões mais precisas recebam maior peso e influência na combinação final. Essa conversão é dada por (21):

$$w_i = \frac{e^{-EQ_i}}{\sum_{j=1}^{3} e^{-EQ_j}} \tag{4.2}$$

Durante o processo de treinamento, o peso atribuído a cada rede neural é atualizado iterativamente, de modo a refletir a contribuição relativa de cada modelo ao longo das épocas. Ao término do treinamento, os valores finais desses pesos são armazenados e posteriormente empregados nas etapas de validação e teste, assegurando consistência na avaliação. Assim, a predição final ($\hat{y}_{final}$) da idade óssea é obtida como uma média ponderada das estimativas individuais, expressa por:

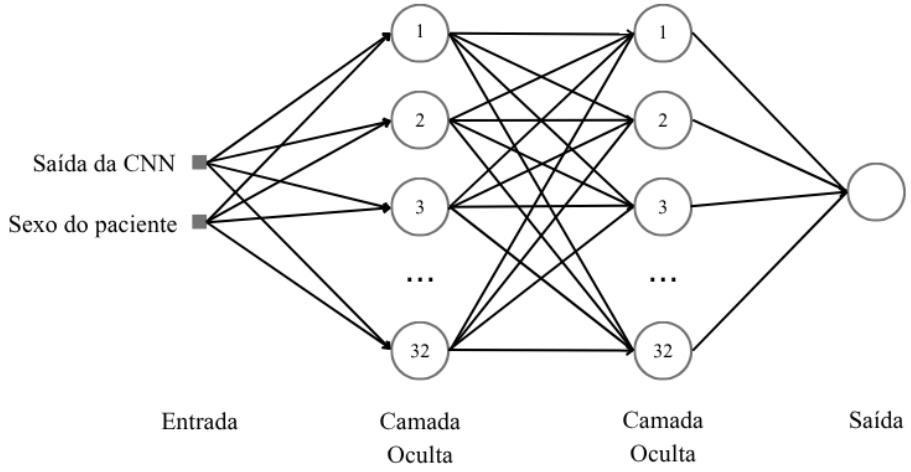
$$\hat{y}_{final} = \sum_{i=1}^{3} w_i \hat{y}_i \tag{4.3}$$

A predição média ponderada das CNNs é então fornecida como entrada para uma rede neural MLP, juntamente com o sexo do paciente. Essa etapa foi desenvolvida com o objetivo de ajustar a predição da idade óssea levando em consideração o sexo

biológico do paciente, motivada por evidências clínicas amplamente documentadas de que o desenvolvimento esquelético apresenta diferenças significativas entre indivíduos do sexo masculino e feminino, o que impacta diretamente na velocidade e padrão de ossificação (Cole *et al.*, 2015).

A arquitetura da MLP foi concebida de maneira simples, composta por duas camadas totalmente conectadas. A camada de entrada recebe dois atributos: a idade óssea predita pela CNN e a codificação numérica do sexo do paciente (0 para feminino e 1 para masculino). Essa entrada é processada por uma camada densa com 32 neurônios e função de ativação ReLU. Em seguida, uma camada linear de saída com um único neurônio produz o valor final refinado da idade óssea, incorporando informações demográficas relevantes sem que seus gradientes influenciem diretamente o treinamento das redes convolucionais. A arquitetura pode ser avaliada na Figura 27.

Figura 27 – Gráfico arquitetônico da MLP



Fonte: Elaborado pela autora.

O processo de otimização foi conduzido com uma taxa de aprendizado inicial de 10^{-3} para a CNN e 10^{-2} para a MLP. O treinamento foi realizado ao longo de 30 épocas, utilizando *batch size* de 64, valor definido em função da capacidade da GPU disponível. Como função de perda, empregou-se o MSE, atribuindo penalidades mais severas a erros de maior magnitude.

5 RESULTADOS

Este capítulo apresenta e discute os resultados obtidos nas duas etapas principais da metodologia proposta. Primeiramente, são analisados os desempenhos do modelo de detecção e segmentação das regiões anatômicas da mão na seção 5.1, incluindo a avaliação quantitativa das métricas obtidas pelo modelo YOLO. Em seguida, na seção 5.2, são apresentados os resultados do modelo de regressão para estimativa da idade óssea, construído a partir das imagens segmentadas. Nesta segunda parte, são detalhadas as métricas de erro obtidas durante o treinamento e validação, além da avaliação final em conjunto de teste independente. Por fim, é realizada uma discussão abrangente sobre os resultados alcançados, comparando com os métodos existentes no estado da arte de modo a contextualizar as contribuições e limitações.

5.1 Modelo de segmentação de regiões

A avaliação da etapa de segmentação das regiões de interesse, implementado com o YOLOv10, foi conduzida de modo a avaliar tanto a capacidade do modelo em localizar corretamente as estruturas de interesse quanto seu potencial prático para uso em um processamento automático de estimativa de idade óssea. Para estimar o desempenho do modelo, foram utilizadas as métricas padronizadas em detecção de objetos, comparando as caixas preditas com as caixas de referência anotadas manualmente.

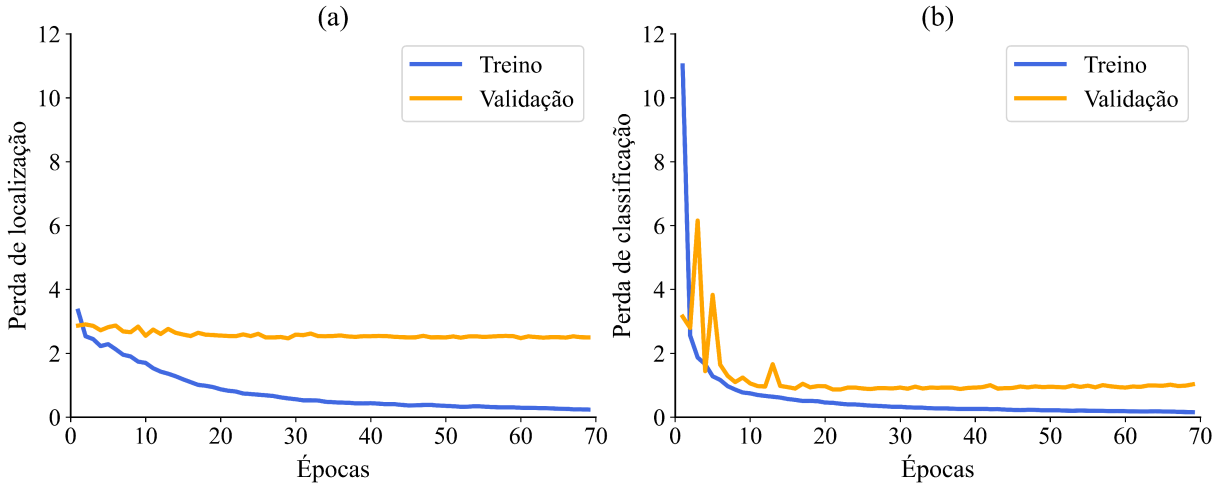
5.1.1 Desempenho em treinamento

Nesta etapa o modelo foi avaliado durante o processo de treinamento e validação, utilizando exclusivamente o subconjunto de dados reservado para essas fases. O objetivo dessa análise inicial foi monitorar a convergência do modelo, verificar a estabilidade das métricas de desempenho e identificar possíveis indícios de sobreajuste. Ao longo das 200 épocas o desempenho do modelo foi monitorado continuamente no conjunto de validação, e com o critério de early stopping para encerrar o treinamento caso não fosse observada melhoria significativa nas métricas de validação após 20 épocas consecutivas, a estabilidade foi atingida na época 69.

O processo de otimização foi conduzido com os hiperparâmetros padrão do YOLOv10 utilizando as funções de perda de localização e de classificação. A perda de localização está associada à precisão dos limites das caixas delimitadoras anotadas e previstas, a sobreposição, tamanho e centro. Já a perda de classificação representa o erro de classificação entre as categorias previstas e as reais. Ambas as perdas devem ser as mais próximas de zero possível.

A Figura 28 apresenta o acompanhamento dessas curvas ao longo das épocas. A queda acentuada da perda de localização nas primeiras épocas, seguida de estabilização em valores próximos de 0.3 no treinamento e 2.5 na validação, demonstra que o modelo aprendeu a localizar adequadamente as estruturas de interesse, com pequenas variações residuais. Para a perda de classificação, a redução para valores próximos de 1 no conjunto de validação demonstra uma boa discriminação entre as classes, com baixo índice de falsos positivos e falsos negativos.

Figura 28 – Evolução das funções de perda para o modelo de segmentação. (a) Perda de localização. (b) Perda de classificação

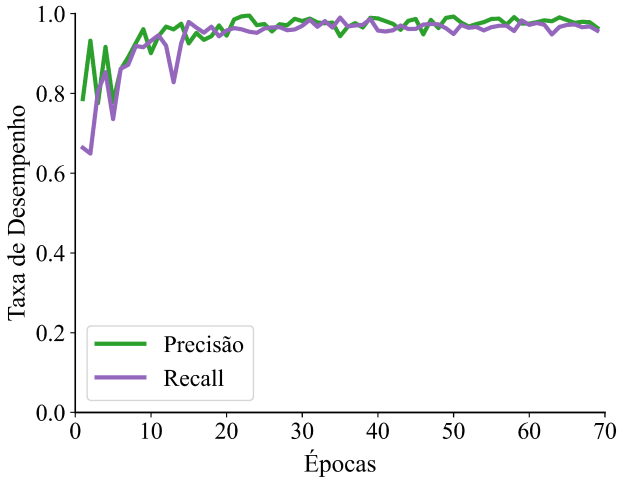


Fonte: Elaborado pela autora.

Além das funções de perda, a Figura 29 apresenta a evolução das métricas de precisão e recall durante o treinamento, calculadas sobre o conjunto de validação. Durante o treinamento, ambas as métricas apresentaram crescimento acentuado nas primeiras épocas, com resultados finais de 97% para precisão e 96% para recall, indicando que o modelo atingiu um equilíbrio satisfatório entre detecções corretas e abrangência das previsões no dataset fornecido.

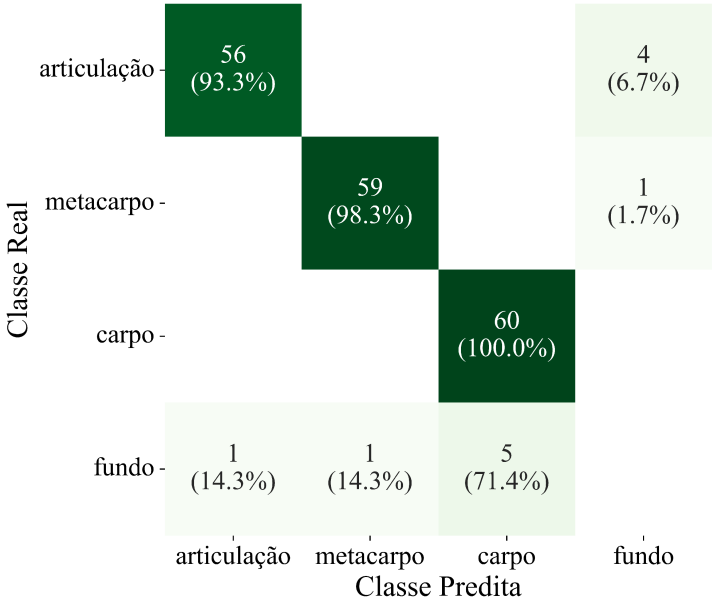
A partir desses indicadores o desempenho do modelo em validação foi analisado para cada classe de interesse, utilizando a matriz de confusão obtida na Figura 30, e calculando as métricas individuais. A matriz mostra que o modelo alcançou desempenho elevado em todas as categorias, com valores predominantes ao longo da diagonal principal, indicando classificações corretas. As ocorrências fora da diagonal representam erros de confusão entre as classes, que, neste caso, são pouco expressivas.

Figura 29 – Evolução das métricas de precisão e recall para o modelo de segmentação



Fonte: Elaborado pela autora.

Figura 30 – Matriz de confusão para o modelo de segmentação em validação



Fonte: Elaborado pela autora.

Para a classe articulação foram registradas 56 classificações corretas e 5 incorretas, sendo 4 delas regiões que não foram identificadas e que o modelo considerou como fundo (falso negativo) e 1 região de fundo que foi classificada como articulação (falso positivo). Assim, a precisão é de 98%, e o recall é de 93%.

Para a classe metacarpo observam-se 59 classificações corretas e 2 incorretas (1 falso positivo e 1 falso negativo). A precisão e o recall para essa classe é de 98%. A classe carpo apresentou 60 predições corretas, com 5 erros de classificação de falsos positivos. Assim, a precisão é 92% e o recall é 100%.

O índice global de acurácia considerando as três classes é de aproximadamente 0.93, o que evidência excelente desempenho geral. Observa-se que os pequenos erros na base

de validação podem ser atribuídas à grande variabilidade introduzida intencionalmente nas etapas de treinamento e validação, com o objetivo de tornar o modelo mais robusto a diferentes padrões de posicionamento e iluminação. Essa diversidade expõe o modelo a cenários mais desafiadores do que aqueles presentes na base original, que em sua maioria apresenta disposições mais padronizadas. Dessa forma, é esperado que o desempenho do modelo na base completa apresente resultados superiores aos observados na validação. A Tabela 3 resume os resultados de validação encontrados.

Tabela 3 – Comparação de resultados de validação para o modelo de segmentação

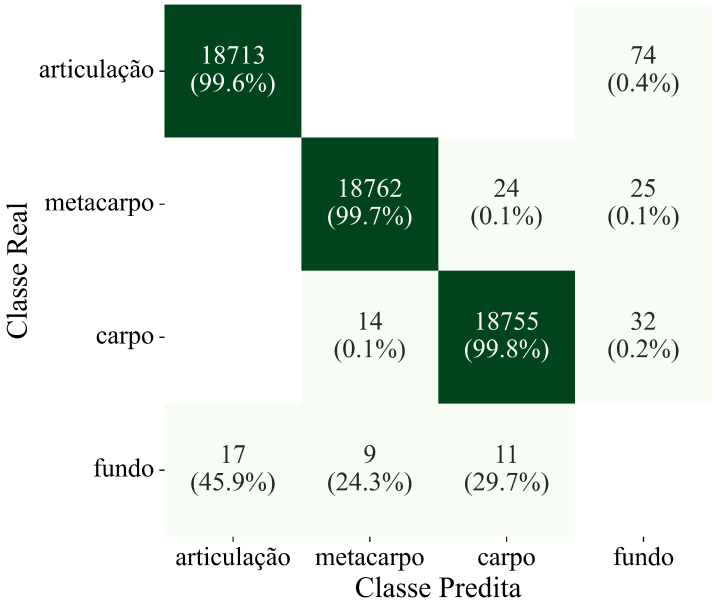
Região	**Precisão**	**Recall**	**Acurácia**
Global	0.97	0.96	0.93
Articulação	0.98	0.93	-
Metacarpo	0.98	0.98	-
Carpo	0.92	1.00	-

5.1.2 Resultados finais

Após a etapa de validação, o modelo foi aplicado sobre o conjunto completo de 18.787 imagens da base de dados. O tempo total de inferência usando apenas CPU foi de 2.188 segundos ($\approx$ 36 minutos), o que corresponde a uma média de aproximadamente 0.11 segundos por imagem, demonstrando a eficiência computacional do modelo YOLOv10s, mesmo considerando o elevado número de detecções necessárias.

A partir da matriz de confusão apresentada na Figura 31, é possível observar que o modelo obteve ótimo desempenho, com boa capacidade de generalização após o treinamento. As três classes de interesse apresentaram taxas de precisão e recall superiores a 99.5%, o que indica um equilíbrio relevante entre as previsões corretas e o número de detecções realizadas para cada categoria. Dessa forma, a acurácia global, definida como a proporção total de classificações corretas, manteve-se igualmente alta, em torno de 99.7%.

Figura 31 – Matriz de confusão para o modelo de segmentação na base total



Fonte: Elaborado pela autora.

O modelo salva automaticamente as imagens com as regiões recortadas diretamente na pasta do projeto, enquanto as imagens correspondentes às detecções ausentes, em um total de 109, foram obtidas manualmente.

5.2 Modelo de detecção da idade óssea

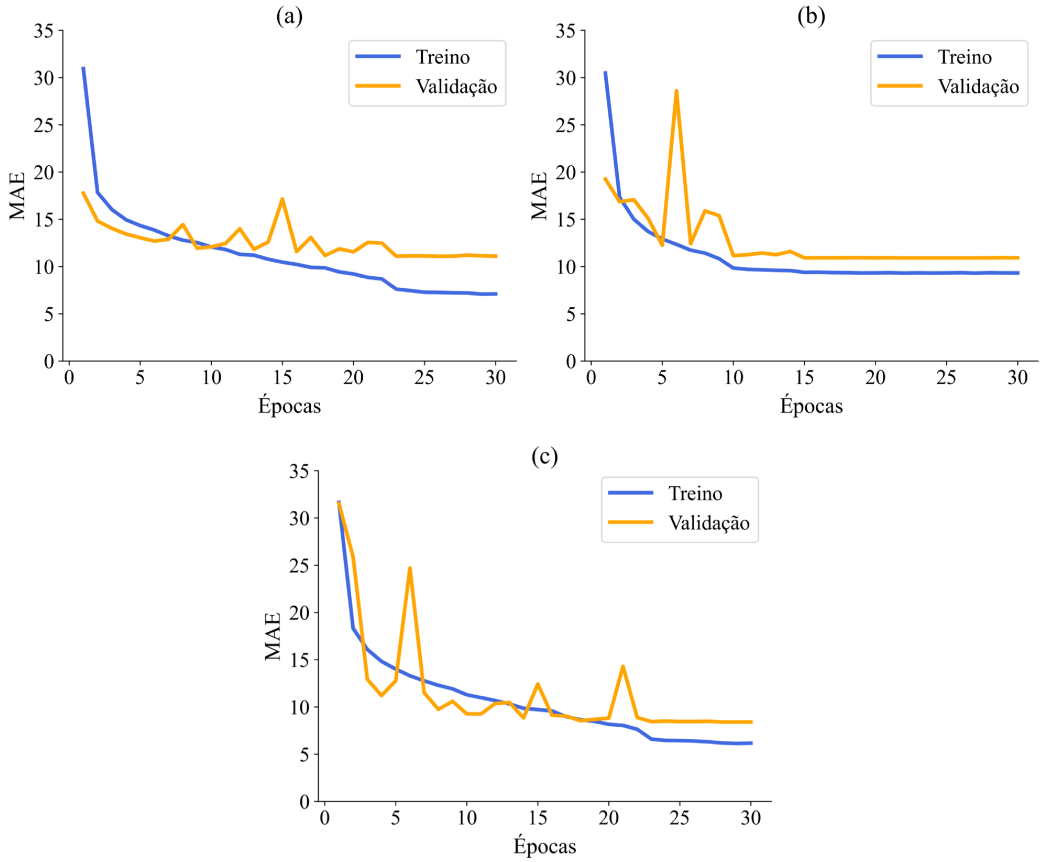
O modelo de estimativa da idade óssea recebe as regiões obtidas na etapa anterior como entrada, extraindo características morfológicas e estruturais presentes nas radiografias. A avaliação do modelo, sendo um problema de regressão, foi conduzida utilizando as métricas de Erro Absoluto Médio (MAE) e Erro Percentual Absoluto Médio (MAPE), selecionadas para fornecer uma visão completa da precisão e do erro do modelo.

5.2.1 Desempenho em treinamento

Esta seção detalha os resultados obtidos durante a fase de treinamento e validação, e o objetivo desta análise é avaliar a capacidade do modelo em aprender os padrões a partir dos dados de treinamento e sua habilidade de generalizar esse aprendizado para dados novos, representados pelo conjunto de validação.

Para aprofundar a compreensão do comportamento do modelo híbrido, será analisado primeiro o desempenho individual de cada uma das três redes convolucionais que processam as diferentes regiões da imagem. A Figura 32 apresenta as curvas do MAE por época, separadamente para os conjuntos de treinamento e validação.

Figura 32 – Curvas de MAE por época para as CNNs de cada região. a) Articulação, b) Metacarpo, c) Carpo



Fonte: Elaborado pela autora.

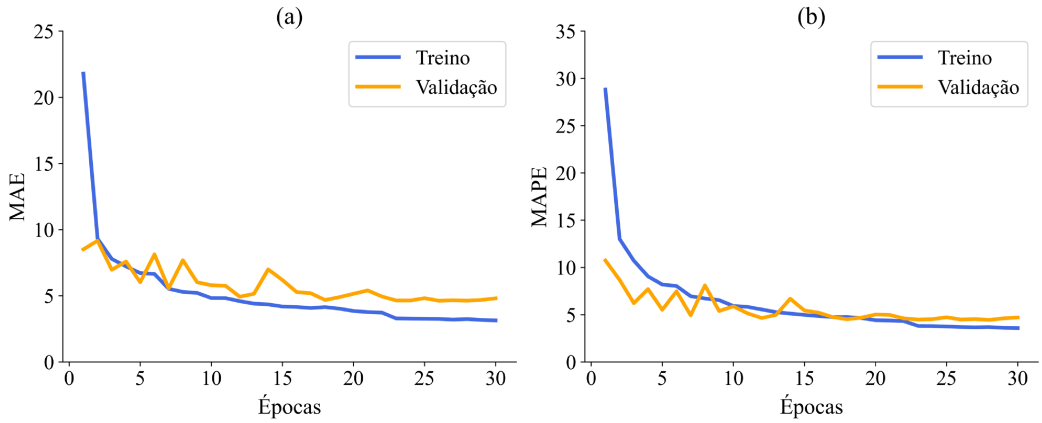
A rede apresentada em (a) é focada nas articulações, e demonstra um comportamento de treinamento robusto e estável. A curva de treinamento exibe uma convergência suave, iniciando com um MAE superior a 30 meses e reduzindo-se consistentemente e estabilizando em 7.62 meses. A curva de validação também apresenta uma queda acentuada inicial, atingindo o ponto de convergência em 11.76 meses. A rede (b) desenvolvida para o metacarpo apresenta uma dinâmica de treinamento diferente, onde a curva de treinamento converge de forma mais acelerada nas épocas iniciais em comparação com a rede anterior, atingindo um platô de erro mínimo mais cedo com um MAE de 9.77 meses. Ambas as redes, apesar de processarem regiões distintas da imagem e exibirem performance de treinamento diferentes, alcançam um desempenho de generalização final próximo, com um MAE de validação de 11.82 meses.

O treinamento da terceira rede em (c) para a região do carpo inicia também com um MAE elevado, e decresce de forma suave e consistente ao longo das 30 épocas. O modelo atingiu um MAE final estabelecendo-se em 6.54 meses com os dados de treinamento. A curva de validação apresenta uma dinâmica com mais instabilidade no começo, porém convergindo para um MAE de validação final estável, em 9.26 meses. Observa-se um

padrão geral de convergência em todas as redes, com o MAE de treinamento diminuindo consistentemente ao longo das épocas e estabilizando-se em limites distintos para cada CNN, indicando que todas as regiões contêm uma quantidade comparável e distinta de informação preditiva.

A Figura 33 apresenta as curvas de convergência do MLP final, responsável pela média das saídas das três CNNs regionais e pela integração da informação do sexo do paciente, gerando a predição definitiva da idade óssea. O gráfico mostra, respectivamente, o MAE e o MAPE ao longo de 30 épocas.

Figura 33 – Curvas de convergência de (a) MAE e (b) MAPE para o modelo MLP final



Fonte: Elaborado pela autora.

As curvas de treinamento demonstram um processo de aprendizado com uma descida acentuada nas épocas iniciais, seguida por uma convergência suave. O MAE de treinamento inicia acima de 20 meses e estabiliza em um valor baixo de 3.57 meses. O MAPE segue um padrão similar, convergindo para 3.73%. Este comportamento indica que o MLP possui capacidade suficiente para se ajustar aos dados de treinamento combinados. As curvas de validação também conseguiram convergir e estabilizar após a época 20, atingindo um platô estável com um MAE de validação final de 4.97 meses, e o MAPE em 5.04%.

O resultado demonstra que o modelo aprendeu a ponderar as saídas das CNNs, e integrar a informação de sexo do paciente como um preditor demográfico chave. A combinação desses fatores permitiu ao MLP final gerar uma predição significativamente mais precisa do que qualquer componente individual. Os resultados ao final das 30 épocas estão consolidados na Tabela 4.

Tabela 4 – Resultados de MAE e MAPE para os modelos CNN e MLP.

Modelo	MAE	MAPE
CNN (a)		
Treino	7.62	7.85%
Validação	11.76	11.40%
CNN (b)		
Treino	9.77	9.89%
Validação	11.82	12.04%
CNN (c)		
Treino	6.54	7.02%
Validação	9.26	10.13%
MLP		
Treino	3.57	3.73%
Validação	4.97	5.04%

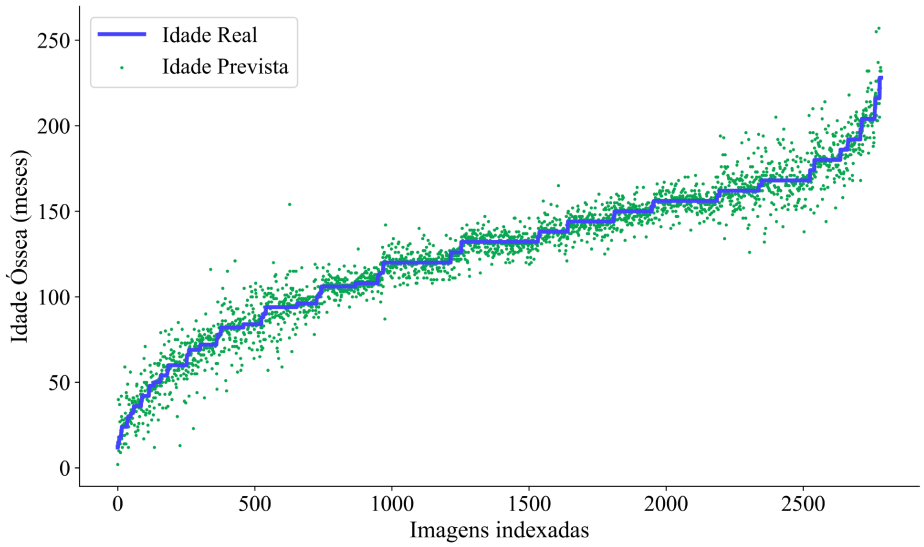
5.2.2 Desempenho em teste

Após a fase de otimização utilizando os conjuntos de treinamento e validação, esta seção apresenta a avaliação final do modelo conduzida no conjunto de teste, um subconjunto de dados isolado que não foi utilizado em nenhuma etapa anterior do desenvolvimento.

A Figura 34 apresenta o gráfico de dispersão das previsões do modelo comparado com os valores reais para todas as amostras do conjunto de teste. Para facilitar a visualização, as imagens no eixo X foram indexadas e ordenadas de acordo com a idade óssea real, representadas pela linha azul do gráfico, e as previsões individuais do modelo são representadas pelos pontos verdes. Em um gráfico de dispersão, a precisão do modelo é inferida pela proximidade dos pontos de dados ao redor da linha que representa a idade real.

Na análise visual deste gráfico observa-se uma nuvem de previsões densa que segue de perto da referência, demonstrando que o modelo aprendeu a relação entre as características da imagem e a idade óssea, mantendo uma alta correlação positiva, especialmente na faixa etária intermediária. No entanto, algumas discrepâncias permanecem entre as idades mais jovens e mais velhas. A principal causa técnica é diretamente visível no histograma de distribuição da base que apresenta sub-representação nas extremidades, que não pode ser totalmente amenizada com técnicas de aumento de dados.

Figura 34 – Gráfico de dispersão das previsões no conjunto de teste



Fonte: Elaborado pela autora.

A avaliação do modelo híbrido final no conjunto de teste resultou nas métricas de desempenho com valores de 6.13 meses para o MAE, e um MAPE de 6.45%. O resultado demonstra que a performance é estável e generalizável, onde os erros são apenas ligeiramente superiores ao observado ao final da convergência no conjunto de validação, e indicando um baixo erro relativo em relação à idade real dos pacientes, validando a arquitetura proposta.

5.3 Discussões

Esta seção tem como objetivo contextualizar os resultados obtidos pelo modelo proposto para avaliação da idade óssea, comparando-os com abordagens relevantes da literatura recente, incluindo os trabalhos fornecidos. Serão analisadas as métricas de desempenho, as escolhas metodológicas, o uso de informações demográficas, a flexibilidade da arquitetura, a composição dos dados de treinamento e as estratégias de validação, identificando as vantagens e limitações do método desenvolvido.

5.3.1 Comparação geral de desempenho

O modelo proposto neste trabalho apresentou um Erro Médio Absoluto (MAE) de 6,13 meses e um Erro Percentual Absoluto Médio (MAPE) de 6,45% no conjunto de validação. Diversos trabalhos prévios reportaram desempenhos semelhantes. (Guo *et al.*, 2022), (Mehta *et al.*, 2021) e (Chu *et al.*, 2018) obtiveram MAEs entre 5,92 e 6,07 meses em seus conjuntos de teste, valores muito próximos ao observado neste estudo. De forma análoga, (Iglovikov *et al.*, 2018) relataram um MAE de 6.10 meses utilizando um conjunto de redes VGG. Esses resultados reforçam que a arquitetura híbrida e regionalizada

aqui proposta é capaz de extrair informações relevantes de maneira eficaz, alcançando desempenho equivalente ao de métodos consolidados na área.

Por outro lado, (Andleeb *et al.*, 2025) reportaram um MAE consideravelmente inferior, de 4,87 meses. Entretanto, tal resultado foi obtido sobre um conjunto de teste clínico específico composto por pacientes com Escoliose Idiopática do Adolescente (EIA). Embora esse modelo também tenha sido treinado sobre a base RSNA, o desempenho superior pode estar relacionado tanto à especificidade do conjunto de teste quanto à arquitetura DenseNet201 empregada.

Alguns estudos, contudo, apresentaram erros substancialmente maiores. (Wibisono *et al.*, 2019) obtiveram MAE de 14,78 meses ao empregar VGG16, com erros ainda maiores para métodos clássicos de aprendizado de máquina. (Zulkifley *et al.*, 2021) alcançaram um MAE de 7,69 meses, enquanto (Sirati-Amsheh *et al.*, 2025) reportaram 9,30 meses com sua abordagem pré-treinada de forma não supervisionada. (Pan *et al.*, 2020), ao empregar conjuntos de regressores sobre features extraídas por Inception-ResNet-V2, reportaram MAE de 8,59 meses. O desempenho superior do modelo aqui apresentado frente a essas abordagens pode ser atribuído a diferenças estruturais na arquitetura ou à estratégia de treinamento adotada.

5.3.2 Comparação de metodologia

A análise comparativa das metodologias evidencia que distintas escolhas arquiteturais e de pré-processamento exercem impacto direto no desempenho final dos modelos. De forma semelhante à proposta de (Iglovikov *et al.*, 2018), o presente estudo também se baseia na segmentação da imagem em regiões específicas de interesse. Os autores treinaram CNNs independentes sobre as regiões carpais e metacarpais/falanges, concluindo que esta última é mais informativa, exceto em fases iniciais do desenvolvimento. Observaram que o conjunto de modelos regionais superava o desempenho das redes individuais, corroborando a relevância da abordagem regional. O método aqui proposto expande essa ideia ao incluir três regiões distintas e realizar a fusão via MLP, em contraste com a simples média ponderada.

(Guo *et al.*, 2022) também exploraram regiões de interesse inspiradas no método Tanner-Whitehouse. O desempenho obtido (MAE de 6,07 meses) é praticamente idêntico ao do presente trabalho, sugerindo que ambas as estratégias são eficazes na combinação das representações regionais. (Zulkifley *et al.*, 2021), por sua vez, recorreram à arquitetura Xception. Apesar da sofisticação da abordagem, o MAE obtido (7,69 meses) pode indicar que a definição manual de regiões, aliada ao MLP, mostrou-se mais eficiente do que essa implementação específica, possivelmente também em função da ausência da variável de sexo. (Chu *et al.*, 2018) abordaram o problema de forma distinta, convertendo a regressão contínua em um conjunto de classificações ordinais (K–1 classes). Essa reformulação

resultou em MAE competitivos (5,98 meses), sugerindo que abordagens ordinais podem oferecer vantagens, embora o presente trabalho mantenha a formulação de regressão direta.

O sexo é um fator clinicamente relevante para a estimativa da idade óssea, e o modelo proposto o incorpora explicitamente como variável de entrada no MLP final. Estudos anteriores também reconheceram sua importância. (Iglovikov *et al.*, 2018) observaram que modelos específicos por sexo podiam superar os modelos mistos, enquanto (Pan *et al.*, 2020) reportaram resultados semelhantes ao treinar redes separadas para cada corte. (Mehta *et al.*, 2021) incluíram a variável de gênero na entrada da rede Inception V3, ao passo que (Zulkifley *et al.*, 2021) optaram por omiti-la por motivos de privacidade, o que pode ter contribuído para o desempenho inferior. A inclusão explícita do sexo, portanto, representa um diferencial metodológico relevante e possivelmente vantajoso em termos preditivos.

Todas as pesquisas utilizaram exclusivamente o dataset RSNA, o que pode limitar a generalização clínica. A variabilidade introduzida utilizando as bases RSNA e RHPE em conjunto representa uma vantagem significativa, aumentando a robustez do modelo. (Andleeb *et al.*, 2025) evidenciaram a importância dessa questão ao testar sua rede em pacientes com EIA, fora da distribuição original da RSNA. Além disso, a metodologia aqui aplicada faz distinção explícita entre os conjuntos de treinamento, validação e teste, reservando este último para avaliação final. Tal prática reduz o risco de otimização indevida de hiperparâmetros sobre o conjunto de teste, evitando estimativas infladas de desempenho, uma limitação observada em estudos que utilizam apenas divisões 70/30 ou 80/20, como (Guo *et al.*, 2022) e (Wibisono *et al.*, 2019).

6 CONCLUSÃO

6.1 Introdução

A aplicação de técnicas de aprendizado de máquina tem impulsionado uma transformação significativa na área médica, especialmente em tarefas diagnósticas baseadas em imagem. Sistemas de diagnóstico computadorizados, potencializados por redes neurais convolucionais, demonstram uma capacidade crescente de identificar padrões complexos, superando em alguns casos a variabilidade humana e oferecendo suporte objetivo à decisão clínica. A avaliação da idade óssea é um exemplo proeminente dessa contribuição, onde a análise automatizada de radiografias de mão e punho auxilia no diagnóstico de distúrbios de crescimento. Este capítulo final tem como objetivo consolidar as conclusões extraídas do presente trabalho, destacando seus resultados, contribuições metodológicas e delineando caminhos para investigações futuras.

6.2 Conclusões Finais

A metodologia proposta neste trabalho constituiu-se de um fluxo de três etapas. Inicialmente, as imagens de radiografias da mão foram submetidas a um estágio de pré-processamento, no qual foram aplicadas técnicas de realce de contraste, com o objetivo de destacar estruturas ósseas relevantes. Em seguida, foi implementado um modelo de detecção de objetos (YOLOv10) para a segmentação automática das regiões de interesse (articulação, metacarpo e carpo). Este modelo de detecção demonstrou um ótimo desempenho, alcançando uma acurácia global de 99,7% na base total. Essa etapa foi essencial para garantir que as imagens utilizadas pelo modelo de regressão representassem de forma precisa as regiões anatômicas relevantes.

A terceira etapa, focada na avaliação da idade óssea, empregou uma arquitetura híbrida projetada para otimizar a extração de características regionais. A metodologia proposta empregou três redes neurais convolucionais (CNNs) customizadas, onde cada rede se especializou em uma região de interesse distinta da radiografia da mão (articulação, metacarpo e carpo). As saídas preditivas de cada CNN foram combinadas através de uma média ponderada, ajustada pelos respectivos erros de validação de cada rede, e concatenadas com a informação demográfica do sexo do paciente. Este vetor de características consolidado serviu como entrada para um Perceptron de Múltiplas Camadas (MLP) final, responsável pela regressão da idade óssea.

A avaliação final do modelo, realizada em um conjunto de teste isolado, confirmou a eficácia e a robustez da abordagem. O modelo alcançou um Erro Médio Absoluto (MAE) de 6,13 meses e um Erro Percentual Absoluto Médio (MAPE) de 6,45%. A proximidade

entre o MAE de teste e o de validação é um indicador forte da capacidade de generalização do modelo. Este desempenho posiciona o modelo proposto de forma competitiva em relação ao estado da arte, atingindo um nível de precisão que não só se equipara a diversas abordagens publicadas, mas também se enquadra dentro da variabilidade interobservador frequentemente aceita na prática clínica.

6.3 Contribuições

Os resultados alcançados evidenciam que este estudo oferece contribuições significativas para o campo do processamento de imagens médicas. A combinação de múltiplas CNNs regionais, especializadas em diferentes áreas anatômicas, com um MLP final para a fusão de características e integração de dados demográficos (sexo), demonstrou ser uma estratégia eficaz, validando a arquitetura de deep learning híbrida e customizada.

Além da arquitetura em si, uma contribuição metodológica significativa foi a utilização de dois conjuntos de dados distintos para o treinamento. Esta abordagem, ainda pouco comum na literatura que predominantemente utiliza apenas a base RSNA, injetou uma variabilidade crucial no processo de treinamento. A exposição do modelo a diferentes equipamentos, protocolos de aquisição e perfis populacionais melhora sua robustez e capacidade de generalização para amostras clínicas do mundo real, que podem divergir da distribuição do conjunto de dados do desafio RSNA.

Adicionalmente, embora modelos de detecção de objetos como o YOLO (You Only Look Once) sejam avançados, sua aplicação específica para a localização automática de ROIs no contexto da avaliação de idade óssea ainda é escassa. A maioria dos trabalhos existentes tende a não empregar técnicas de segmentação específicas ou a utilizar modelos de segmentação distintos, geralmente baseados em abordagens convencionais. O uso do YOLO representa uma abordagem promissora e relevante para o avanço na automatização do fluxo de análise, com alta fidelidade e baixo custo de anotação, contribuindo para tornar o processo mais eficiente e independente de intervenção manual.

As estratégias adotadas para o pré-processamento, segmentação e extração de regiões demonstraram ser eficientes, de implementação simples e com baixo custo computacional, o que as torna adequadas para diferentes aplicações baseadas em imagens radiográficas digitalizadas. A solução proposta foi integralmente desenvolvida neste trabalho, apresentando flexibilidade para ser treinada com distintos conjuntos de dados e demandando um tempo de treinamento compatível com aplicações práticas, e um rápido tempo de inferência.

6.4 Trabalhos Futuros

Embora os resultados obtidos sejam promissores, este trabalho abre diversas oportunidades para investigações futuras e aprimoramentos. As seguintes direções são propostas:

- Ampliar o conjunto de dados utilizado no treinamento, especialmente nas faixas etárias com menor representatividade, a fim de reduzir o desbalanceamento e aprimorar a capacidade de generalização do modelo. Essa expansão pode incluir tanto a coleta de novos exames quanto o uso de data augmentation avançado ou geração de imagens sintéticas por meio de redes generativas.

- Explorar novas técnicas de pré-processamento de imagens, com o objetivo de otimizar a segmentação das estruturas ósseas da mão e do punho. Métodos baseados em filtragem adaptativa, equalização local de contraste e normalização de intensidade podem contribuir para uma extração mais precisa das regiões de interesse.

- Avaliar a aplicabilidade do modelo em diferentes bases externas, oriundas de hospitais ou bancos de imagens públicos, a fim de verificar sua capacidade de generalização e robustez frente a variações de protocolo de aquisição e qualidade das radiografias.

- Aprimorar a arquitetura da rede neural convolucional (CNN), realizando ajustes finos de Transfer Learning com modelos pré-treinados em bases médicas de raios X, de modo a acelerar o treinamento e melhorar a extração de características relevantes.

- Investigar diferentes configurações arquiteturais, como o aumento do número de camadas convolucionais e da resolução das imagens de entrada, bem como a execução em plataformas computacionais mais potentes, visando avaliar o impacto desses fatores sobre a acurácia e o tempo de processamento.

REFERÊNCIAS

ACHARYA, T.; RAY, A. K. **Image processing: principles and applications**. [*S.l.: s.n.*]: John Wiley & Sons, 2005.

AMARI, S.-i. Backpropagation and stochastic gradient descent method. **Neurocomputing**, Elsevier, v. 5, n. 4-5, 1993. DOI: https://doi.org/10.1016/0925-2312(93)90006-O.

ANDLEEB, I. *et al.* Automatic evaluation of bone age using hand radiographs and pancorporal radiographs in adolescent idiopathic scoliosis. **Diagnostics**, MDPI, v. 15, n. 4, 2025. DOI: https://doi.org/10.3390/diagnostics15040452.

BALDOVINO, R. G. *et al.* Comprehensive analysis on ultralytics-supported yolo models for detection and recognition of large office objects for indoor navigation. **Procedia Computer Science**, Elsevier, v. 246, 2024. DOI: https://doi.org/10.1016/j.procs.2024.09.158.

BENGIO, Y. Practical recommendations for gradient-based training of deep architectures. *In*: **Neural networks: Tricks of the trade: Second edition**. [*S.l.: s.n.*]: Springer, 2012. p. 437–478. DOI: https://doi.org/10.1007/978-3-642-35289-$8_2$6.

BISHOP, C. M.; NASRABADI, N. M. **Pattern recognition and machine learning**. [*S.l.: s.n.*]: Springer, 2006. v. 4.

BUCKLER, J. How to make the most of bone ages. **Archives of disease in childhood**, BMJ Publishing Group, v. 58, n. 10, 1983. DOI: https://doi.org/10.1136/adc.58.10.761.

BÜKEN, B. *et al.* Is the assessment of bone age by the greulich–pyle method reliable at forensic age estimation for turkish children? **Forensic science international**, Elsevier, v. 173, n. 2-3, 2007. DOI: https://doi.org/10.1016/j.forsciint.2007.02.023.

BULL, R. *et al.* Bone age assessment: a large scale comparison of the greulich and pyle, and tanner and whitehouse (tw2) methods. **Archives of disease in childhood**, BMJ Publishing Group Ltd, v. 81, n. 2, 1999. DOI: https://doi.org/10.1136/adc.81.2.172.

CAVALLO, F. *et al.* Evaluation of bone age in children: a mini-review. **Frontiers in Pediatrics**, Frontiers Media SA, v. 9, 2021. DOI: https://doi.org/10.3389/fped.2021.580314.

CHU, M. *et al.* Bone age assessment based on two-stage deep neural networks. *In*: IEEE. **2018 Digital Image Computing: Techniques and Applications (DICTA)**. [*S.l.: s.n.*], 2018. p. 1–6. DOI: https://doi.org/10.1109/DICTA.2018.8615764.

COLE, T. J. *et al.* Ethnic and sex differences in skeletal maturation among the birth to twenty cohort in south africa. **Archives of disease in childhood**, BMJ Publishing Group Ltd, v. 100, n. 2, 2015. DOI: https://doi.org/10.1136/archdischild-2014-306399.

CUNNINGHAM, C. *et al.* Bone development. **Developmental Juvenile Osteology**, Elsevier, 2016. DOI: https://doi.org/10.1016/B978-0-12-382106-5.00003-7.

DIGHE, P. C.; GURU, S. K. Survey on image resizing techniques. **International Journal of Science and Research (IJSR)**, v. 3, n. 12, 2014.

ESCOBAR, M. *et al.* Hand pose estimation for pediatric bone age assessment. *In*: SPRINGER. **International conference on medical image computing and computer-assisted intervention**. [*S.l.: s.n.*], 2019. p. 531–539. DOI: https://doi.org/10.1007/978-3-030-32226-$7_5$9.

GILSANZ, V.; RATIB, O. **Hand bone age: a digital atlas of skeletal maturity**. [*S.l.: s.n.*]: Springer, 2005.

GONZALES, R. C.; WINTZ, P. **Digital image processing**. [*S.l.: s.n.*]: Addison-Wesley Longman Publishing Co., Inc., 1987.

GOODFELLOW, I. *et al.* **Deep learning**. [*S.l.: s.n.*]: MIT press Cambridge, 2016. v. 1.

GREULICH, W.; PYLE, S. **Radiographic Atlas of Skeletal Development of the Hand and Wrist**. Stanford University Press, 1959. ISBN 9780804703987. Disponível em: https://books.google.com.br/books?id=olezJFYxM6oC.

GUO, L. *et al.* Bone age assessment based on deep convolutional features and fast extreme learning machine algorithm. **Frontiers in Energy Research**, Frontiers Media SA, v. 9, 2022. DOI: https://doi.org/10.3389/fenrg.2021.813650.

HAYKIN, S. **Neural Networks and Learning Machines**. Prentice Hall, 2009. (Neural networks and learning machines, v. 10). ISBN 9780131471399. Disponível em: https://books.google.com.br/books?id=K7P36lKzI_QC.

HOU, Y. *et al.* The state-of-the-art review on applications of intrusive sensing, image processing techniques, and machine learning methods in pavement monitoring and analysis. **Engineering**, Elsevier, v. 7, n. 6, 2021. DOI: https://doi.org/10.1016/j.eng.2020.07.030.

IGLOVIKOV, V. I. *et al.* Paediatric bone age assessment using deep convolutional neural networks. *In*: SPRINGER. **Deep Learning in Medical Image Analysis and Multimodal Learning for Clinical Decision Support: 4th International Workshop, DLMIA 2018, and 8th International Workshop, ML-CDS 2018, Held in Conjunction with MICCAI 2018, Granada, Spain, September 20, 2018, Proceedings 4**. [*S.l.: s.n.*], 2018. p. 300–308. DOI: https://doi.org/10.1007/978-3-030-00889-$5_3$4.

JIMÉNEZ-CASTELLANOS, J. *et al.* Skeletal maturation of wrist and hand ossification centers in normal spanish boys and girls: a study using the greulich-pyle method. **Cells Tissues Organs**, S. Karger AG Basel, Switzerland, v. 155, n. 3, 1996. DOI: https://doi.org/10.1159/000147806.

JOCHER, G.; QIU, J. **Ultralytics YOLO11**. 2024. Disponível em: https://github.com/ultralytics/ultralytics.

KANDEL, I.; CASTELLI, M. The effect of batch size on the generalizability of the convolutional neural networks on a histopathology dataset. **ICT express**, Elsevier, v. 6, n. 4, 2020. DOI: https://doi.org/10.1016/j.icte.2020.04.010.

KHAN, K.; ELAYAPPEN, A. S. Bone growth estimation using radiology (greulich–pyle and tanner–whitehouse methods). **Handbook of growth and growth monitoring in health and disease**, Springer, 2012. DOI: https://doi.org/10.1007/978-1-4419-1795-$9_1$76.

KIM, S. Y.; YANG, S. W. Assessment of bone age: A comparison of the greulich pyle method to the tanner whitehouse method. **Endocrinology and Metabolism**, Korean Endocrine Society, v. 13, n. 2, 1998.

KOWO-NYAKOKO, F. *et al.* Evaluation of two methods of bone age assessment in peripubertal children in zimbabwe. **Bone**, Elsevier, v. 170, 2023. DOI: https://doi.org/10.1016/j.bone.2023.116725.

LITJENS, G. *et al.* A survey on deep learning in medical image analysis. **Medical image analysis**, Elsevier, v. 42, 2017. DOI: https://doi.org/10.1016/j.media.2017.07.005.

MALINA, R. M. *et al.* **Growth, maturation, and physical activity**. [*S.l.: s.n.*]: Human kinetics, 2004.

MAW, J. *et al.* Hand anatomy. **British Journal of Hospital Medicine**, MA Healthcare London, v. 77, n. 3, 2016. DOI: https://doi.org/10.12968/hmed.2016.77.3.C34.

MEHTA, C. *et al.* Deep learning framework for automatic bone age assessment. *In*: IEEE. **2021 43rd Annual International Conference of the IEEE Engineering in Medicine & Biology Society (EMBC)**. [*S.l.: s.n.*], 2021. p. 3093–3096. DOI: https://doi.org/10.1109/EMBC46164.2021.9629650.

MELMED, S. *et al.* **Williams textbook of endocrinology E-Book**. [*S.l.: s.n.*]: Elsevier Health Sciences, 2015.

MUGHAL, A. M. *et al.* Bone age assessment methods: a critical review. **Pakistan journal of medical sciences**, Professional Medical Publications, v. 30, n. 1, 2014. DOI: https://doi.org/10.12669/pjms.301.4295.

MUMUNI, A.; MUMUNI, F. Data augmentation: A comprehensive survey of modern approaches. **Array**, Elsevier, v. 16, 2022. DOI: https://doi.org/10.1016/j.array.2022.100258.

NIELSEN, M. A. **Neural networks and deep learning**. [*S.l.: s.n.*]: Determination press San Francisco, CA, USA, 2015. v. 25.

PAN, X. *et al.* Fully automated bone age assessment on large-scale hand x-ray dataset. **International journal of biomedical imaging**, Hindawi, 2020. DOI: https://doi.org/10.1155/2020/8460493.

PAN, Y. Different types of neural networks and applications: Evidence from feedforward, convolutional and recurrent neural networks. **Highlights in Science, Engineering and Technology**, v. 85, 03 2024. DOI: https://doi.org/10.54097/6rn1wd81.

PARSANIA, P. *et al.* A review: Image interpolation techniques for image scaling. **International Journal of Innovative Research in Computer and Communication Engineering**, v. 2, n. 12, 2014. DOI: https://doi.org/10.15680/IJIRCCE.2014.0212024.

PINCHI, V. *et al.* Skeletal age estimation for forensic purposes: A comparison of gp, tw2 and tw3 methods on an italian sample. **Forensic science international**, Elsevier, v. 238, 2014. DOI: https://doi.org/10.1016/j.forsciint.2014.02.030.

PRATT, W. K. **Digital image processing: PIKS Scientific inside**. [*S.l.: s.n.*]: Wiley Online Library, 2007. v. 4.

QUEIROZ, J. E. R. de; GOMES, H. M. Introdução ao processamento digital de imagens. **Rita**, v. 13, n. 2, 2006.

RAIAAN, M. A. K. *et al.* A systematic review of hyperparameter optimization techniques in convolutional neural networks. **Decision Analytics Journal**, Elsevier, v. 11, 2024. DOI: https://doi.org/10.1016/j.dajour.2024.100470.

RASAMOELINA, A. D. *et al.* A review of activation function for artificial neural network. *In*: IEEE. **2020 IEEE 18th world symposium on applied machine intelligence and informatics (SAMI)**. [*S.l.: s.n.*], 2020. p. 281–286. DOI: https://doi.org/10.1109/SAMI48414.2020.9108717.

REDMON, J. *et al.* You only look once: Unified, real-time object detection. *In*: **Proceedings of the IEEE conference on computer vision and pattern recognition**. [*S.l.: s.n.*], 2016. p. 779–788. DOI: https://doi.org/10.1109/CVPR.2016.91.

RIPLEY, B. D. **Pattern recognition and neural networks**. [*S.l.: s.n.*]: Cambridge university press, 2007.

RSNA. 2017. Disponível em: https://www.rsna.org/education/ai-resources-and-training/ai-image-challenge/rsna-pediatric-bone-age-challenge-2017.

SANCTIS, V. D. *et al.* Hand x-ray in pediatric endocrinology: Skeletal age assessment and beyond. **Indian journal of endocrinology and metabolism**, Medknow, v. 18, n. Suppl 1, 2014. DOI: https://doi.org/10.4103/2230-8210.145076.

SATOH, M. Bone age: assessment methods and clinical applications. **Clinical Pediatric Endocrinology**, The Japanese Society for Pediatric Endocrinology, v. 24, n. 4, 2015. DOI: https://doi.org/10.1297/cpe.24.143.

SHAH, N. *et al.* Comparison of bone age assessments by gruelich-pyle, gilsanz-ratib, and tanner whitehouse methods in healthy indian children. **Indian Journal of Endocrinology and Metabolism**, Medknow, v. 25, n. 3, 2021. DOI: https://doi.org/10.4103/ijem.IJEM$_8$26$_2$0.

SHOME, S. K.; VADALI, S. R. K. Enhancement of diabetic retinopathy imagery using contrast limited adaptive histogram equalization. **International Journal of Computer Science and Information Technologies**, Citeseer, v. 2, n. 6, 2011.

SIRATI-AMSHEH, M. *et al.* Ae-bonet: A deep learning method for pediatric bone age estimation using an unsupervised pre-trained model. **Journal of Biomedical Physics & Engineering**, v. 15, n. 3, 2025.

TAN, L.; JIANG, J. **Digital signal processing: fundamentals and applications**. [*S.l.: s.n.*]: Academic press, 2018.

TANNER, J. *et al.* Prediction of adult height from height, bone age, and occurrence of menarche, at ages 4 to 16 with allowance for midparent height. **Archives of disease in childhood**, BMJ Publishing Group Ltd, v. 50, n. 1, 1975. DOI: https://doi.org/10.1136/adc.50.1.14.

TERVEN, J. *et al.* A comprehensive review of yolo architectures in computer vision: From yolov1 to yolov8 and yolo-nas. **Machine learning and knowledge extraction**, MDPI, v. 5, n. 4, 2023. DOI: https://doi.org/10.3390/make5040083.

THODBERG, H. H. *et al.* The bonexpert method for automated determination of skeletal maturity. **IEEE transactions on medical imaging**, IEEE, v. 28, n. 1, 2008. DOI: https://doi.org/10.1109/TMI.2008.926067.

TOET, A.; WU, T. Efficient contrast enhancement through log-power histogram modification. **Journal of Electronic Imaging**, Society of Photo-Optical Instrumentation Engineers, v. 23, n. 6, 2014. DOI: https://doi.org/10.1117/1.JEI.23.6.063017.

VAQUERO, D. *et al.* A survey of image retargeting techniques. *In*: SPIE. **Applications of digital image processing XXXIII**. [*S.l.: s.n.*], 2010. v. 7798, p. 328–342. DOI: https://doi.org/10.1117/12.862419.

WANG, Z. *et al.* A comprehensive survey on data augmentation. 2024.

WIBISONO, A. *et al.* Deep learning and classic machine learning approach for automatic bone age assessment. *In*: IEEE. **2019 4th Asia-Pacific Conference on Intelligent Robot Systems (ACIRS)**. [*S.l.: s.n.*], 2019. p. 235–240. DOI: https://doi.org/10.1109/ACIRS.2019.8935965.

ZAFAR, A. M. *et al.* An appraisal of greulich-pyle atlas for skeletal age assessment in pakistan. **JPMA. The Journal of the Pakistan Medical Association**, v. 60, n. 7, 2010.

ZHANG, C. *et al.* Understanding deep learning requires rethinking generalization. 2016.

ZHANG, S.-Y. *et al.* Standards of tw3 skeletal maturity for chinese children. **Annals of human biology**, Taylor & Francis, v. 35, n. 3, 2008. DOI: https://doi.org/10.1080/03014460801953781.

ZULKIFLEY, M. A. *et al.* Intelligent bone age assessment: an automated system to detect a bone growth problem using convolutional neural networks with attention mechanism. **Diagnostics**, MDPI, v. 11, n. 5, 2021. DOI: https://doi.org/10.3390/diagnostics11050765.