UNIVERSIDADE FEDERAL DO MARANHÃO

UNIVERSIDADE FEDERAL DO PIAUÍ

Doutorado em Ciência da Computação Associação UFMA/UFPI

Pablo de Abreu Vieira

Modelos Preditivos e Generativos para Análise Automatizada de Radiografias da Coluna Lombo-Sacra e Pododáctilos

Orientador: Profº. Drº. Romuere Rodrigues Veloso e Silva

Co-orientador: Profº. Drº. Mano Joseph Mathew

Teresina - PI
Novembro, 2024

Pablo de Abreu Vieira

Modelos Preditivos e Generativos para Análise Automatizada de Radiografias da Coluna Lombo-Sacra e Pododáctilos

TESE DE DOUTORADO

Tese apresentada como requisito parcial para obtenção do título de Doutor em Ciência da Computação, ao Doutorado em Ciência da Computação, Associação UFMA/UFPI.

Orientador: Profº. Drº. Romuere Rodrigues Veloso e Silva
Co-orientador: Profº. Drº. Mano Joseph Mathew

Teresina - PI
Novembro, 2024

Universidade Federal do Piauí

Biblioteca Comunitária Jornalista Carlos Castello Branco

Divisão de Representação da Informação

V658m Vieira, Pablo de Abreu.

Modelos preditivos e generativos para análise automatizada de radiografias da coluna lombo-sacra e pododactilos / Pablo de Abreu Vieira. – – Teresina, 2024.

149 f. : il.

Tese (Doutorado) – Universidade Federal do Piauí, Universidade Federal do Maranhão, Programa de Pós-Graduação em Ciência da Computação, 2024.

"Orientador: Profº. Drº. Romuere Rodrigues Veloso e Silva"

1. Diagnóstico assistido por computador. 2. Inteligência artificial generativa. 3. *Transformers*. 4. Raio X. I. Silva, Romuere Rodrigues Veloso e. II. Título.

CDD 006.3

Elaborada por Fabíola Nunes Brasilino CRB 3/ 1014

Pablo de Abreu Vieira

Modelos Preditivos e Generativos para Análise Automatizada de Radiografias da Coluna Lombo-Sacra e Pododáctilos

A presente Tese de Doutorado foi avaliada e aprovada por banca examinadora composta pelos seguintes membros:

Profº. Drº. Romuere Rodrigues Veloso e Silva
Orientador
Universidade Federal do Piauí

Profº. Drº. Profº. Drº. Mano Joseph Mathew
Co-orientador
École d'Ingénieurs Généraliste du Numérique

Profª. Drª. Andrea Gomes Campos
Examinadora Externa
Universidade Federal de Ouro Preto

Profª. Drª. Fátima Nelsizeuma Sombra de Medeiros
Examinadora Externa
Universidade Federal do Ceará

Profº. Drº. Antonio Oseas de Carvalho Filho
Examinador Interno
Universidade Federal do Piauí

Profº. Drº. Flávio Henrique Duarte de Araújo
Examinador Interno
Universidade Federal do Piauí

Certificamos que esta é a versão original e final da Tese de Doutorado que foi julgada adequada para obtenção do título de Doutora em Ciência da Computaçã

gov.br Documento assinado digitalmente
ROMUERE RODRIGUES VELOSO E SILVA
Data: 23/12/2024 14:52:49-0300
Verifique em https://validar.iti.gov.br

Profº. Drº. Romuere Rodrigues Veloso e Silva
Orientador

gov.br Documento assinado digitalmente
IVAN SARAIVA SILVA
Data: 24/12/2024 07:47:54-0300
Verifique em https://validar.iti.gov.br

Profº. Drº. Ivan Saraiva Silva
Coordenador

Teresina - PI, 04 de Novembro de 2024

À minha esposa e filhos, que são a fonte constante de inspiração e amor em minha vida. Vocês são o combustível que me impulsiona a seguir em frente e a buscar sempre o melhor.

Agradecimentos

Agradecemos ao Cosmos, de onde tudo se originou — dos átomos às galáxias, dos corpos celestes ao nosso planeta, que abriga a mais deslumbrante natureza e nós, os seres sapiens, que contemplamos e buscamos compreender a vastidão do universo.

Aos meus pais, Maurício Vieira e Maria Abreu, que me deram o dom da vida e, com amor incondicional, moldaram meu caráter e orientaram minha busca pelo melhor. Sua sabedoria e dedicação foram faróis que iluminaram meu caminho.

Ao meu irmão, Maurício Filho, que, ao longo de muitas aventuras e desafios, sempre esteve ao meu lado como um companheiro leal e encorajador. Sua presença constante e apoio foram essenciais em cada passo dessa jornada.

Aos meus padrinhos, Manoel Lima e Mairan Vieira, que sempre deram suporte fundamental para minha educação. Em especial, ao meu padrinho, Manoel Lima, cuja memória e influência continuam a inspirar-me e guiar-me, mesmo após seu falecimento. Agradeço profundamente por tudo o que ele fez por mim, e sua presença será eternamente lembrada e apreciada.

À minha esposa, Luzana Brasileiro, que tem sido uma fonte constante de inspiração e apoio ao longo de minha trajetória profissional. Sua presença é uma peça fundamental nesta conquista acadêmica, e sou profundamente grato por sua paciência e compreensão durante os momentos difíceis, quando a tentação de desistir era grande. Sua força e amor incondicional foram essenciais para que eu chegasse até aqui.

A meus filhos, Felipe e Rafael, que me proporcionaram um amor transformador e profundo, o qual me inspira a ser uma pessoa melhor do que eu poderia sonhar. O amor que sinto por vocês é a força motriz por trás de cada conquista e de cada passo que dou. Vocês tornam minha vida mais significativa e me incentivam a buscar sempre o melhor.

Ao meu orientador, Professor Dr. Romuere Silva, expresso minha sincera gratidão pela oportunidade de realizar este trabalho. Seus ensinamentos foram fundamentais para o desenvolvimento de meu mestrado e, agora, para meu doutorado. Agradeço por sua paciência e pelo esclarecimento das minhas dúvidas, desde meus primeiros passos na pós-graduação até esta etapa final, celebrando nossas vitórias no mestrado e o avanço para o doutorado.

To my co-advisor, Mano Mathew, who welcomed me with open arms during my sandwich Ph.D. program in Paris. The experience of living in another country, far from home, was challenging, but your guidance and support were essential in helping me overcome difficulties and enriching my academic journey. I am deeply grateful for your

support and dedication during this period.

À minha antiga coorientadora, Professora Drª. Deborah Magalhães, agradeço imensamente por seu empenho e apoio no inicio dessa jornada. Seu bom humor e disposição para esclarecer dúvidas e revisar meus trabalhos foram cruciais.

Aos professores que compõem a banca, agradeço pelo tempo e esforço dedicados à avaliação deste trabalho. Suas contribuições e insights são de grande valor e enriqueceram significativamente este estudo.

Aos professores das disciplinas do doutorado, expresso minha gratidão por suas valiosas contribuições:

- Professor Pedro Alcântara, pela dedicação e excelência na disciplina de Engenharia de Software, que foi fundamental para o aprofundamento dos meus conhecimentos na área.
- Professor André Soares, pelos ensinamentos na disciplina de Redes, que ampliaram minha compreensão e habilidades nesse campo crucial.
- Professor Rodrigo Veras, coordenador do doutorado e professor da disciplina de Projeto e Análise de Algoritmos, cuja orientação e conhecimento foram essenciais para a realização de projetos desafiadores e complexos.
- Professor Alselmo Paiva, pela condução do Workshop, que proporcionou uma experiência prática e enriquecedora, complementando minha formação acadêmica de maneira significativa.

À Coordenação de Aperfeiçoamento de Pessoal de Nível Superior - Brasil (CAPES), Código de Financiamento 88881.846250/2023-01, expresso minha profunda gratidão pelo apoio financeiro que possibilitou a realização do meu doutorado sanduíche. Sem esse suporte, a oportunidade de viver e estudar em Paris não teria sido viável.

Também agradeço ao Conselho Nacional de Desenvolvimento Científico e Tecnológico (CNPq), Código de Financiamento 311289/2022-3, cujo apoio foi fundamental para a realização deste trabalho

Aos meus antigos colegas de trabalho:

- Da SDU-Sudeste: Adriana, Lina, Jefferson, Ámerico, Fonteneles, Azevedo e seu Zé, que foram meus companheiros de trabalho por muitos e muitos anos. Continuamos a nos apoiar mutuamente como uma verdadeira família.
- Da UFPI: Patricia Medyna, Patricia Vieira, Alcilene Dalila, Francisco Imperes, Denis Carvalho, Ismael Leal, Márcio e demais professores, por ter me mostrado que o meio acadêmico era o lugar onde eu sempre iria querer está.

- Da Maida.Health: Luís Henrique, Lucas Lopes, Luís Guilherme, Lucas Bezerra e Dyogo, onde aprendi muito do que apliquei nesta tese. A experiência adquirida nesse ambiente foi valiosa para meu desenvolvimento.
- Da SSP-PI: Felipe Venceslau, Joaquim Carvalho, Neto Araújo, Rogério Paulo, José Alberto, Teodoro, Ramon e Yan, que me receberam de braços abertos e contribuíram significativamente para minha trajetória.
- Ao corpo técnico da Inatel, onde aplicarei muito do que aprendi neste doutorado, agradeço pelo suporte e pela oportunidade de continuar crescendo profissionalmente.

Às escolas onde passei:

- Colégio Sagrado da Conceição: Onde iniciei meus passos na educação, do Jardim à terceira série. Agradeço à Tia Núbia, diretora da escola, por seu apoio; à Tia Virgínia, quem me alfabetizou; e à Tia Neide, que conviveu muito comigo. Agradeço também aos meus amigos Bruno Almeida, Amauri, Vinícius, Rafael, Luisão, Fabiano, Leonel, Jéssica, Nayara, Adriana, Amanda e todos os outros que fizeram parte dessa fase.
- Santa Joana D'Arc: Minha segunda escola, onde estudei a quarta série e pude entender que existiam outros mundos escolares.
- Instituto Nossa Senhora do Amparo: Onde estudei da quinta à oitava série. Agradeço à finada diretora Simirinda e à minha professora do coração Irene, que foram fundamentais para minha formação. Ao professor Chico Carlos, que me fez entender a importância da gramática portuguesa; ao professor Peterson, que me fez amar a história; à professora Vânia, que me despertou a paixão pelas ciências; os professores Chagas e Barbozinha que me fizeram adorar as aulas de matemática; e aos meus amigos do coração André Rezende, Emanuel Rezende, Adriana, Werton, Rafaela, Romuere Silva, Danilo, Débora, Dayane, Tácio Carmargo, entre muitos outros.
- Méroto D'Martone: Onde estudei o ensino médio. Agradeço ao diretor Expedito e ao professor Sidney, que me fez amar biologia. Agradeço também aos meus amigos de sempre: Felipe Ramos, um irmão; Renato; José Neto; Leonardo Monteiro; Bruna Werclose; e Emanuel Resende.

Ao Instituto de Ensino Superior de Teresina, agradeço pela oportunidade de conhecer e desenvolver minha área de atuação em computação, agradeço meus amigos Mauro Sergio, Selemérico, Wesley, Mario Tales, Mauricio e todos os outros.

À Universidade Santo Agostinho, onde me apaixonei por Redes de Computadores, agradeço ao meu orientador professor MsC. Ricardo Queiroz e aos meus amigos de LPI Janiel, Ciro, Atevaldo e todos os outros.

À UFPI, que me proporcionou tanto trabalho quanto a possibilidade de fazer meu mestrado e agora o doutorado.

A todos os meus amigos, em especial ao Nonato Sales e ao Thiago José, que nos momentos mais difíceis e turbulentos não mediram esforços para me ajudar, tanto no mestrado quanto neste doutorado. Desde a correção de um simples bug ao entendimento de conceitos complexos, foram meus maiores parceiros nesta caminhada. Agradeço também a Daniel e Leonel Feitosa, que me deram um suporte inestimável neste doutorado.

Aos meus amigos da EFREi, Christiano que me ajudou em praticamente quase tudo em minha estadia em Paris e hoje é quase um irmão, ao Iago que sempre me mostrava o lado leve de viver em outro pais sozinho e ao Marciel que foi um grande companheiro que esteve ao meu lado no pior momento de minha estadia na França, obrigado a todos.

Por fim, quero estender meus agradecimentos a todos, mesmo àqueles não mencionados especificamente aqui. Recebam meus sinceros agradecimentos, do fundo do coração, pela contribuição e apoio ao longo desta jornada.

“Não se coloque dentro de uma forma, se adapte e construa sua própria, e deixa-a expandir, como a água. Se colocarmos a água num copo, ela se torna o copo; se você colocar água numa garrafa ela se torna a garrafa. A água pode fluir ou pode colidir. Seja água, meu amigo.”

Bruce Lee.

Resumo

A radiografia é uma ferramenta essencial na medicina moderna para a visualização interna do corpo humano. Essa técnica, amplamente utilizada para examinar diversas patologias, como doenças ósseas e alterações em tecidos moles, fornece imagens detalhadas com base na absorção diferencial dos raios-X por tecidos de diferentes densidades. No entanto, a análise de radiografias apresenta desafios significativos, como a identificação de anomalias sutis e a interpretação precisa das imagens, muitas vezes dificultada pela variabilidade dos exames e pela sobrecarga de trabalho dos radiologistas. Esta tese propõe uma abordagem inovadora para otimizar a triagem de exames e a geração de laudos médicos preliminares a partir de radiografias, com foco nas radiografias da coluna lombo-sacra e dos pododáctilos. A metodologia desenvolvida combina técnicas de inteligência artificial preditiva e generativa, utilizando arquiteturas avançadas de redes neurais convolucionais como Inception-V3, VGG e ResNet50, aliadas a modelos generativos baseados em *Transformers*, para enfrentar os desafios associados à análise radiográfica. O objetivo principal é melhorar a eficiência na triagem de exames e fornecer laudos médicos automatizados como suporte à decisão clínica, oferecendo uma segunda opinião detalhada para auxiliar no diagnóstico tanto na coluna lombo-sacra quanto nos pododáctilos. A pesquisa utilizou dois conjuntos de dados radiográficos de lombo-sacra composto por 44.540 e de pododáctilos composto por 16.710. As imagens foram coletadas de hospitais e clinicas de todo território brasileiro, abrangendo uma variedade de cenários clínicos. O desenvolvimento de modelos preditivos, baseado em aprendizado profundo, focou na triagem e filtragem de radiografias, identificando exames que requerem maior atenção e otimizando a eficiência do processo. Além disso, foram desenvolvidos modelos generativos, usando técnicas de aumento de dados e *fine-tuning*, para a geração automática de laudos médicos preliminares, complementando o trabalho dos radiologistas com diagnósticos e descrições detalhadas a partir das imagens analisadas. Os resultados obtidos na metodologia de classificação de patologias de lombo-sacra, utilizando a técnica de ensemble, demonstraram alta precisão, com valores de 0,941 (Acurácia), 0,874 ($kappa$), 0,983 (Precisão), 0,952 (F1-Score), 0,947 (AUC), 0,972 (Especificidade) e 0,983 (Sensibilidade). Já na geração automática de laudos para radiografias de lombo-sacra, as métricas alcançadas foram: 0,612 (BLEU-1), 0,552 (BLEU-2), 0,507 (BLEU-3), 0,470 (BLEU-4), 0,471 (METEOR) e 0,633 (ROUGE-L). Para os laudos gerados a partir de radiografias de pododáctilos, os resultados foram: 0,516 (BLEU-1), 0,432 (BLEU-2), 0,386 (BLEU-3), 0,370 (BLEU-4), 0,414 (METEOR) e 0,364 (ROUGE-L). Esses resultados destacam a eficácia do método desenvolvido, que demonstrou potencial para melhorar a precisão dos diagnósticos, otimizar o processo de triagem e oferecer uma solução prática e eficiente para a geração automática de laudos médicos, contribuindo significativamente para o avanço da prática clínica moderna.

Palavras-chave: X-rays, Lumbosacral Spine, Pododactyls, Convolutional Neural Networks (CNNs), Transformadores, Generative Artificial Intelligence, Computer-Aided Diagnosis (CAD)

Abstract

Radiography is an essential tool in modern medicine for visualizing the internal structures of the human body. This technique, widely used to examine various pathologies, such as bone diseases and soft tissue abnormalities, provides detailed images based on the differential absorption of X-rays by tissues of different densities. However, the analysis of radiographs presents significant challenges, including the identification of subtle anomalies and the accurate interpretation of images, often hindered by variability in exams and the heavy workload of radiologists. This thesis proposes an innovative approach to optimize the triage of exams and the generation of preliminary medical reports from radiographs, focusing on lumbosacral and pododactyl radiographs. The developed methodology combines predictive and generative artificial intelligence techniques, utilizing advanced convolutional neural network architectures such as Inception-V3, VGG, and ResNet50, alongside generative models based on Transformers, to address the challenges associated with radiographic analysis. The main objective is to improve the efficiency of exam triage and provide automated medical reports as clinical decision support, offering a detailed second opinion to assist in the diagnosis of both the lumbosacral spine and pododactyls. The research used two radiographic datasets: a lumbosacral dataset comprising 44,540 images and a pododactyl dataset comprising 16,710 images. The images were collected from hospitals and clinics across Brazil, encompassing a variety of clinical scenarios. The development of predictive models, based on deep learning, focused on triaging and filtering radiographs, identifying exams that require further attention, and optimizing the efficiency of the process. Additionally, generative models were developed using data augmentation and fine-tuning techniques for the automatic generation of preliminary medical reports, complementing the radiologists' work with detailed diagnoses and descriptions from the analyzed images. The results obtained for the classification of lumbosacral pathologies using an ensemble technique demonstrated high accuracy, with values of 0.941 (Accuracy), 0.874 (κ), 0.983 (Precision), 0.952 (F1-Score), 0.947 (AUC), 0.972 (Specificity), and 0.983 (Sensitivity). For the automatic generation of reports for lumbosacral radiographs, the achieved metrics were: 0.612 (BLEU-1), 0.552 (BLEU-2), 0.507 (BLEU-3), 0.470 (BLEU-4), 0.471 (METEOR), and 0.633 (ROUGE-L). For the reports generated from pododactyl radiographs, the results were: 0.516 (BLEU-1), 0.432 (BLEU-2), 0.386 (BLEU-3), 0.370 (BLEU-4), 0.414 (METEOR), and 0.364 (ROUGE-L). These results highlight the effectiveness of the developed method, which demonstrated potential to improve diagnostic accuracy, optimize the triage process, and offer a practical and efficient solution for the automatic generation of medical reports, contributing significantly to the advancement of modern clinical practice.

Key-words: X-rays, Lumbosacral Spine, Pododactyls, Convolutional Neural Networks

(CNNs), Transformers, Generative Artificial Intelligence, Computer-Aided Diagnosis (CAD)

Lista de ilustrações

Lista de tabelas

Lista de abreviaturas e siglas

AD	Aumento de Dados
AHE	*Adaptive Histogram Equalization*
ANNs	*Artificial Neural Network*
AP	Aprendizado Profundo
BERT	*Bidirectional Encoder Representations from Transformers*
BLEU	*Bilingual Evaluation Understudy*
CAD	*Computer-Aided Diagnosis*
CDF	*Cumulative Distribution Function*
CLAHE	*Contrast Limited Adaptive Histogram Equalization*
CTC	Camadas Totalmente Conectadas
CV	Coluna Vertebral
DAC	Diagnóstico Auxiliado por Computador
DICOM	*Digital Imaging and Communications in Medicine*
FMM	*Fast Marching Method*
FN	Falso Negativo
FP	Falso Positivo
GANs	*Generative Adversarial Networks*
Grad-CAM	*Gradient-weighted Class Activation Mapping*
VGG	*Visual Geometry Group*
GPT	*Generative Pre-trained Transformer*
IV3	*Inception-V3*
IA	Inteligência Artificial
MC	Matriz de Confusão

METEOR	*Metric for Evaluation of Translation with Explicit ORdering*
MLP	*Multilayer Perceptron*
NLP	*Natural Language Processing*
OCR	*Optical Character Recognition*
PM	Perceptron de Multicamadas
RIR	*Interpretation of Report by Radiologist*
RF	*Random Forests*
RNCs	Redes Neurais Convolucionais
RNAs	Redes Neurais Artificiais
RN50	*ResNet50*
ROC	*Receiver Operating Characteristic*
SOTA	*State-of-the-Art*
SVM	*Support Vector Machines*
VP	Verdadeiro Positivo
VN	Verdadeiro Negativo
TA	Transferência de Aprendizado

Sumário

1 Introdução

A radiografia tem evoluído e se consolidado como uma das principais ferramentas de diagnóstico por imagem na medicina moderna, sendo amplamente utilizada para a detecção de patologias em diferentes partes do corpo explorando o princípio de que tecidos com diferentes densidades absorvem raios-X de maneiras distintas. Esta técnica resulta em contrastes variados na imagem radiográfica, onde estruturas densas, como ossos, absorvem mais radiação e aparecem de forma proeminente, enquanto estruturas menos densas, como os pulmões e seus gases, são visíveis devido à sua menor absorção (HILL et al., 2014; SEMA, 2019; NR, 2011; VIEIRA et al., 2021). Essa capacidade de diferenciar tecidos com base em sua densidade torna as radiografias uma ferramenta essencial para a detecção de patologias do esqueleto e doenças em tecidos moles.

A radiografia é uma tecnologia consolidada, amplamente acessível e econômica, sendo o método de imagem médica mais utilizado na maioria dos centros de saúde. Este exame, quando realizado por técnicos experientes, é minimamente invasivo e possui um tempo de execução relativamente curto, permitindo a visualização de diversas partes do corpo em um único procedimento. Esses fatores contribuem para a frequente utilização dos raios-X no diagnóstico de várias condições médicas, reforçando seu papel fundamental na prática clínica diária (REUMATOLOGIA, 2011).

As radiografias são amplamente utilizadas em exames clínicos para avaliar condições médicas variadas, incluindo radiografias de tórax (BENNOUR et al., 2024), pododáctilos (CALVO-WRIGHT et al., 2023), coluna lombo-sacra (WENG et al., 2019), seios da face (KIM et al., 2019), abdômen (SANGNARK et al., 2024), pelve (LEE et al., 2020), crânio (LEE et al., 2024), articulações (SALEEM et al., 2020) e extremidades (STEVENS, 2020), dentre outras. Elas também são fundamentais na detecção de problemas em órgãos internos, músculos e dentes. Essa versatilidade demonstra a importância das radiografias na prática clínica moderna, permitindo a visualização detalhada de estruturas anatômicas e a identificação de diversas patologias (HILL et al., 2014; SEMA, 2019; NR, 2011; WANG et al., 2017).

Apesar das vantagens da radiografia, a geração de laudos médicos de qualidade por especialistas a partir dessas imagens não é trivial. Esse processo exige profissionais especializados, cuja carga de trabalho e condições estressantes podem levar a resultados imprecisos. Nesse contexto, há uma crescente necessidade de métodos e ferramentas que auxiliem os especialistas em suas análises, dando origem aos sistemas de Diagnóstico Assistido por Computador (*Computer-Aided Diagnosis*, CAD) (VIEIRA et al., 2021;

ESMAIL; EL-DIN; A., 2020; Zeng et al., 2020; KONG et al., 2024). Esses sistemas são projetados para apoiar a interpretação das imagens radiográficas, melhorando a precisão diagnóstica e reduzindo o impacto da fadiga e do estresse nos profissionais de saúde.

Os sistemas CAD estão sendo cada vez mais adotados para auxiliar na triagem e fornecer uma segunda opinião aos especialistas, utilizando técnicas de inteligência artificial (IA) em seu desenvolvimento. Exames de radiografias da coluna lombo-sacra e dos pododáctilos são especialmente promissores para esses estudos, pois fornecem imagens detalhadas dessas regiões. Isso possibilita a interpretação por modelos de visão computacional e *machine learning*, tanto na classificação dos exames quanto na geração automática de laudos, oferecendo suporte valioso aos especialistas e melhorando a precisão e a eficiência diagnóstica (VIEIRA et al., 2021; VIEIRA et al., 2023b; VIEIRA et al., 2024).

A IA tem proporcionado avanços na análise de exames de raios-X, especialmente no desenvolvimento de sistemas CAD. Em particular, a IA preditiva, por meio de algoritmos de *deep learning*, permite a análise de grandes volumes de dados radiográficos, identificando automaticamente padrões e anomalias que podem não ser detectados devido à carga de trabalho dos especialistas. Esses modelos são eficazes na triagem de imagens, classificando-as conforme a probabilidade de patologias específicas e priorizando os casos que requerem atenção imediata (VIEIRA et al., 2023b).

Por outro lado, a IA generativa é capaz de criar novos dados como por exemplo textos com base em padrões aprendidos, fornecendo uma segunda opinião e informações adicionais que podem auxiliar na interpretação dos exames. Modelos generativos podem sugerir diagnósticos e gerar laudos médicos iniciais, complementando o trabalho dos radiologistas e servindo como uma etapa inicial no processo de elaboração de laudos (PAVLOPOULOS et al., 2022; YI; WALIA; BABYN, 2019).

A interpretação de radiografias, especialmente na detecção de anomalias sutis ou raras, é um desafio devido à complexidade e à variabilidade das imagens médicas. Para enfrentar esse problema, o presente estudo propõe o desenvolvimento de um sistema CAD que combina modelos de IA preditiva para o processo de triagem com modelos de IA generativa para a automatização da geração de laudos médicos preliminares em radiografias da coluna lombo-sacra e pododáctilos.

Este estudo, aborda a aplicação de IA para auxiliar especialistas na análise de exames de coluna lombo-sacra e pododáctilos, integrando modelos de IA preditivas para a filtragem e triagem de exames, e modelos de IA generativa para o suporte na elaboração de laudos médicos. A IA preditiva visa otimizar a identificação e classificação de patologias, priorizando os casos mais críticos, enquanto a IA generativa oferece uma segunda opinião, auxiliando na tomada de decisão clínica. Com essa abordagem, esperamos aprimorar tanto a precisão e a eficiência diagnóstica, quanto reduzir a carga

de trabalho e o risco de erros dos profissionais de saúde, promovendo um ambiente clínico mais seguro e eficaz.

1.1 Objetivo Geral

O objetivo geral desta é desenvolver um sistema CAD que combine modelos de IA preditivas, para a triagem de exames, e modelos de IA generativas, para a geração automática de laudos médicos preliminares a partir de radiografias da coluna lombo-sacra e dos pododáctilos, com o intuito de auxiliar médicos no processo diagnóstico.

1.2 Objetivos Específicos

Para alcançar o objetivo geral deste trabalho, faz-se necessário atingir os seguintes objetivos específicos:

- Desenvolver um modelo de IA preditiva para triagem de radiografias de coluna lombo-sacra, otimizando a identificação de exames que necessitam de atenção prioritária.
- Criar e treinar um modelo de IA generativa para auxiliar na geração de laudos médicos (preliminares) descritivos e diagnósticos, a partir de radiografias da coluna lombo-sacra e dos pododáctilos. Esses laudos servirão como suporte à análise dos especialistas, oferecendo uma segunda opinião que contribui para a precisão e eficiência dos diagnósticos médicos.
- Avaliar o desempenho dos modelos desenvolvidos, comparando-os com métodos existentes e utilizando métricas específicas para validar a precisão, confiabilidade e eficiência na triagem e na qualidade dos laudos.

1.3 Contribuições

As principais contribuições do método proposto nesta tese são descritas a seguir:

1. Desenvolvimento de um modelo preditivo inovador para a triagem de radiografias, baseado em CNNs e utilizando arquiteturas como VGG16, VGG19 e ResNet50. O modelo aplica uma estratégia de *Ensemble* com seleção de limiares de confiança para ambas as incidências (frontal e lateral), incorporando técnicas de *fine-tuning* e aumento de dados.
2. Implementação de um modelo generativo inovador para geração automática de laudos médicos preliminares, baseado em *Transformers*, para interpretar

características visuais extraídas pela CNN Inception-V3. Foram aplicadas técnicas de *fine-tuning* e aumento de dados para gerar laudos a partir de radiografias de lombo-sacra e pododáctilos. Os conjuntos de dados utilizados incluíram 16.710 radiografias de pododáctilos e 44.540 de lombo-sacra, com o objetivo de auxiliar radiologistas fornecendo descrições detalhadas e diagnósticos precisos.

1.4 Organização do Trabalho

Os demais capítulos deste trabalho foram organizados da seguinte forma:

- O Capítulo 2 discute os conceitos fundamentais desta pesquisa, abordando temas como exames de coluna lombo-sacra e pododáctilos, pré-processamento de imagens digitais, *machine learning, deep learning*, redes neurais convolucionais, inteligência artificial generativa e *transformers*.
- O Capítulo 3 apresenta os trabalhos relacionados, com foco na classificação de sistemas CAD para a classificação de anomalias em radiografias e geração automática de laudos médicos.
- O Capítulo 4 detalha o método proposto para a classificação de exames da coluna lombo-sacra no processo de triagem, bem como os resultados e as conclusões obtidas.
- O Capítulo 5 aborda o método proposto para a geração automática de laudos médicos preliminares dos exames de lombo-sacra e pododáctilos, além dos resultados e conclusões correspondentes.
- Finalmente, o Capítulo 6 apresenta as considerações finais, destacando as contribuições do trabalho, as limitações encontradas e as sugestões para trabalhos futuros.

2 Fundamentação Teórica

Este capítulo detalha os tópicos essenciais para a compreensão das técnicas empregadas na elaboração dos métodos propostos. Nas seções subsequentes, discutimos conceitos relacionados aos exames de coluna lombo-sacra e pododáctilos, pré-processamento de imagens digitais, *machine learning, deep learning*, redes neurais convolucionais, transferência de aprendizado e ajuste fino, IA generativa em *Transformers*, aumento de dados e as métricas de desempenho utilizadas para validar os resultados experimentais.

2.1 Exames de Lombo-Sacra

A anatomia da coluna vertebral humana compreende um conjunto de 33 vértebras, das quais 24 são separadas por discos intervertebrais. Essa estrutura desempenha funções essenciais de suporte estrutural e mobilidade, além de servir como conduto para abrigar a medula espinhal. A integridade da coluna é mantida por uma rede intricada de ligamentos e articulações, conferindo-lhe flexibilidade. Anatomicamente, a coluna é dividida em cinco regiões distintas: cervical, torácica, lombar, sacral e coccígea. Distúrbios como osteoartrite, escoliose, espinha bífida, espondiloartrite, hiperlordose, osteófitos e redução do espaço discal, entre outros, podem afetar essa estrutura, resultando em implicações clínicas significativas (DRAKE; VOGL; MITCHELL, 2019; WINESKI, 2024).

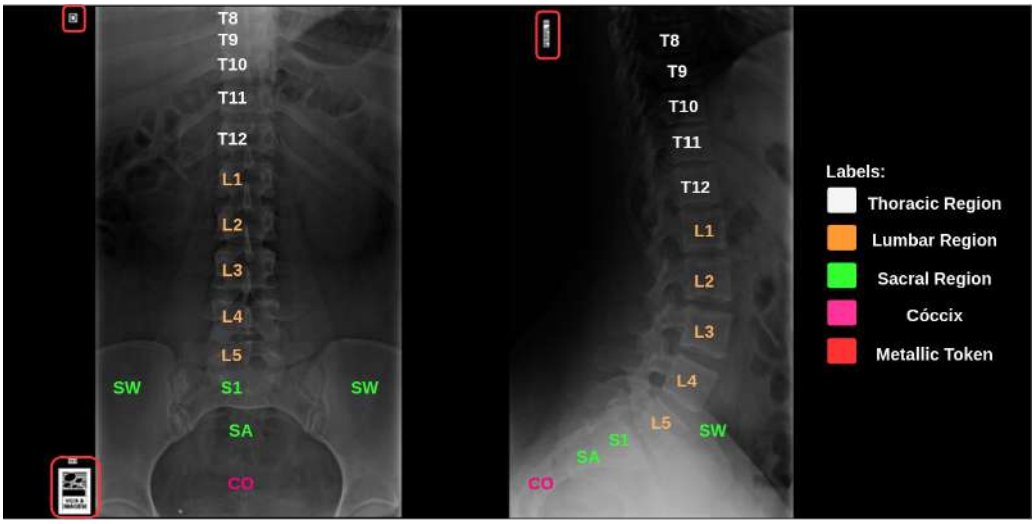
Esses distúrbios têm o potencial de induzir desconforto significativo e, em casos graves, podem levar à imobilização do paciente (ANOUAR, 2019; O.A., 2019). O tratamento dessas condições é essencial, considerando que afetam aproximadamente 80% da população mundial ao longo da vida (TREVOR; ROSAIRE; DRISCOLL, 2022). Estudos indicam que cerca de um terço das pessoas com essas condições necessita de reabilitação (CIEZA et al., 2020). Nos Estados Unidos, os custos com cuidados lombares atingiram USD 134,5 bilhões, com um crescimento anual de 6,7% entre 1996 e 2016. Além disso, exames desnecessários de ressonância magnética da coluna lombar geraram custos adicionais de cerca de USD 300 milhões por ano (BUCHBINDER et al., 2020).

A investigação de problemas relacionados à coluna vertebral frequentemente recorre a exames de imagem, sendo a radiografia uma opção fundamental. Esse método é amplamente utilizado devido ao seu custo acessível, menor invasividade e rapidez na execução (VIEIRA et al., 2021).

O exame de lombo-sacra foca especificamente na região lombar e no sacro da coluna vertebral, abrangendo as cinco vértebras lombares (L1 a L5) e o sacro, que é formado pela fusão de cinco vértebras sacrais fundidas. Esta área é importante para a

mobilidade e suporte estrutural do corpo, sendo um ponto comum de dor e lesões devido à sua carga biomecânica (VIEIRA et al., 2023b). A Figura 1 ilustra um exemplo desse exame, mostrando claramente as regiões lombar e sacral. A Figura 1 também apresenta visualizações frontais e laterais do exame, permitindo uma visão detalhada das estruturas vertebrais e suas possíveis patologias.

Figura 1 – Exemplo de exame de lombo-sacra.



Fonte: o próprio autor.

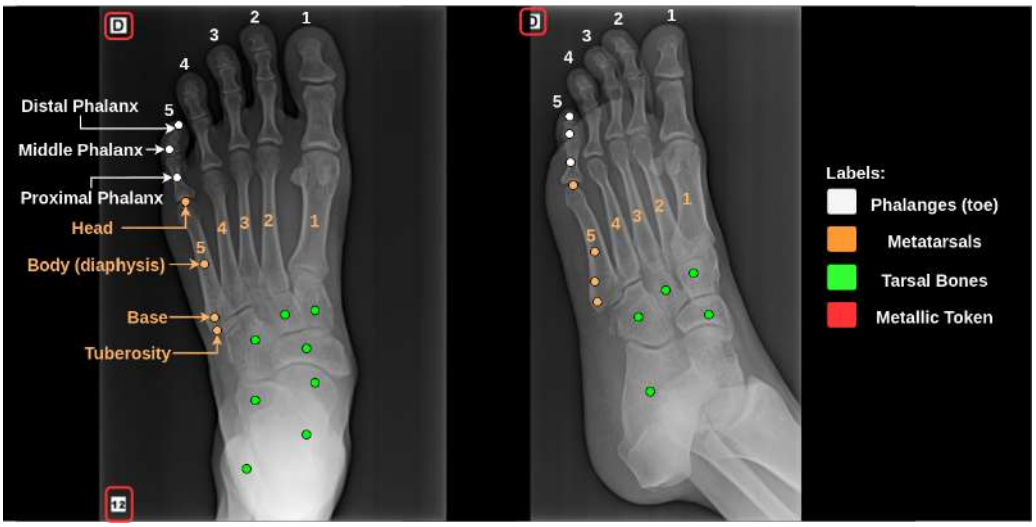
2.2 Exames de Pododáctilos

A anatomia dos pés (pododáctilos) humanos compreende um conjunto de ossos e articulações responsáveis pela estrutura e mobilidade dos pés. Essa estrutura sustenta o peso corporal, facilita a locomoção e desempenha um papel essencial na distribuição de forças e cargas durante o movimento (GEBO, 1988; TOMASSONI; TRAINI; AMENTA, 2014). Anatomicamente, os pododáctilos são divididos em regiões distintas, como o tarso[1], metatarso[2] e falanges[3], desempenhando um papel fundamental na postura e locomoção (SALTZMAN; NAWOCZENSKI, 1995; CAVANAGH et al., 1997; MATTHEWS, 1998). Além de suas funções biomecânicas, diversas condições clínicas podem afetar os pododáctilos, como fascite plantar (TROJIAN; TUCKER, 2019), esporão do calcanhar (BERGMANN, 1990), pé chato (BOERUM; SANGEORZAN, 2003), osteoartrite (RODDY; MENZ, 2018), halux valgo (DESCHAMPS et al., 2010), descalcificação (PENSEC et al., 2004) e lesões traumáticas (GRUSHKY et al., 2021) impactando significativamente a qualidade de vida dos indivíduos.

[1] O tarso é a parte posterior do pé e contém sete ossos que formam o tornozelo e parte do arco do pé.
[2] O metatarso é a parte intermediária do pé e é composto por cinco ossos longos que conectam o tarso às falanges.
[3] As falanges são os ossos dos dedos dos pés, com cada dedo possuindo três falanges, exceto o hálux, que possui duas.

O exame de pododáctilos, focado na análise dos pés, é essencial para avaliar essa complexa estrutura mecânica, que consiste em 26 ossos, 33 articulações (20 das quais são ativamente articuladas), além de mais de cem músculos, tendões e ligamentos. As principais articulações do pé incluem o tornozelo, a articulação subtalar e as articulações interfalângicas. Estudos antropométricos, como aquele que avaliou 1.197 homens caucasianos adultos norte-americanos com idade média de 35,5 anos, indicam que o comprimento médio do pé masculino é de 26,3 cm, com um desvio padrão de 1,2 cm (HAWES; SOVAK, 1994). Esses dados são fundamentais para entender as variações anatômicas e biomecânicas no exame de pododáctilos. A Figura 2 ilustra um exame de pododáctilos, apresentando visualizações frontal e lateral, permitindo uma observação detalhada dessas estruturas e suas possíveis patologias.

Figura 2 – Exemplo de exame de pododáctilos.



Fonte: o próprio autor.

Embora a radiografia seja benéfica para a identificação de doenças e a elaboração de relatórios médicos, a interpretação dessas imagens exige expertise especializada. A natureza repetitiva e detalhada da tarefa pode levar à fadiga, resultando em erros potenciais. Nesse contexto, sistemas CAD, utilizando técnicas de machine learning, estão sendo pesquisados para auxiliar na detecção e diagnóstico. Esses sistemas visam simplificar o trabalho dos especialistas, fornecendo uma segunda opinião, o que pode aumentar a eficiência e reduzir a ocorrência de erros, melhorando a qualidade do diagnóstico (ESMAIL; EL-DIN; A., 2020; Zeng et al., 2020; VIEIRA et al., 2021; VIEIRA et al., 2023a).

2.3 Processamento Digital de Imagens

O pré-processamento de imagens digitais envolve a aplicação de algoritmos à imagem original com o objetivo de melhorar sua qualidade para os processos

subsequentes. Esse processo é essencial para suprimir distorções indesejadas, como ruídos, ou realçar características importantes da imagem. Modelos de visão computacional, *machine learning* ou IA podem se beneficiar significativamente dos dados aprimorados, aumentando a precisão e eficácia nas análises (GONZALEZ, 2011; PEDRINI; SCHWARTZ, 2008; VIEIRA et al., 2021).

Diversos algoritmos contribuem para a metodologia de processamento de imagens. Embora as abordagens possam diferir em suas etapas ou ferramentas, elas normalmente seguem o fluxo proposto por Gonzalez e Woods (GONZALEZ, 2011), que inclui as seguintes fases. Sendo à primeira Aquisição das Imagens, as imagens são capturadas por um dispositivo ou sensor e transformadas em um formato processável. Este processo pode introduzir falhas na imagem resultante, devido a condições de iluminação ou características do dispositivo de captura. Segunda Pré-processamento, o objetivo é melhorar a qualidade da imagem utilizando técnicas para reduzir o ruído, realçar o contraste e suavizar certas estruturas da imagem. Terceira a segmentação visa localizar e isolar áreas de interesse na imagem, separando os diferentes objetos presentes. Quarta Representação e Descrição, conhecida como extração de características, essa etapa envolve a obtenção de informações que podem ser usadas para discriminar classes de objetos. Quinta e última Reconhecimento e Interpretação, o reconhecimento atribui identificadores aos objetos da imagem, enquanto a interpretação atribui significado ao conjunto de objetos reconhecidos.

2.4 Equalização de Histograma

Nas radiografias, a variação na densidade dos ossos e tecidos moles resulta em transições suaves entre diferentes estruturas anatômicas, o que pode dificultar a detecção de anomalias (VIEIRA et al., 2021). Esse problema de contraste ocorre devido à forma como os raios-X interagem com diferentes materiais no corpo humano como os ossos mais densos absorvem mais raios-X e aparecem mais brancos nas radiografias, enquanto os tecidos moles, como os pulmões, absorvem menos raios-X e aparecem mais escuros. No entanto, as transições graduais entre áreas de diferentes densidades podem resultar em bordas pouco definidas, tornando a identificação de detalhes críticos mais desafiadora (SCHEIBEL et al., 2009).

Para mitigar esse problema de contraste, uma técnica amplamente utilizada é a *Adaptive Histogram Equalization* (AHE)(KETCHAM ROGER W. LOWE, 1974). Uma variante mais eficaz dessa técnica é a *Contrast Limited Adaptive Histogram Equalization* (CLAHE)(PIZER et al., 1990), que aprimora o contraste da imagem ao calcular histogramas em pequenas regiões da imagem e ajustá-los individualmente. Embora o AHE possa amplificar ruídos, o CLAHE foi desenvolvido para limitar esse efeito, garantindo uma equalização controlada do contraste sem amplificação excessiva de

ruídos.

O CLAHE divide a imagem em blocos menores e aplica um limite de corte (*Clip Limit*) para evitar a superamplificação de ruídos. O histograma de cada região é ajustado para limitar picos altos, redistribuindo o excesso de forma uniforme entre os *bins* do histograma (PIZER et al., 1990). Esse método garante que o contraste seja aprimorado de maneira eficaz, sem a introdução de artefatos significativos.

O processo matemático do CLAHE pode ser descrito em três passos principais começando pela divisão da imagem em pequenas regiões (ou blocos), pelo cálculo do histograma para cada uma das regiões e aplicação do limite de contraste. Se um *bin*[4] do histograma exceder o valor predefinido (*Clip Limit*), o excesso é redistribuído uniformemente para os outros bins, evitando o super-realce de áreas da imagem.

A função de distribuição acumulativa (*Cumulative Distribution Function* - CDF) é então utilizada para remapear os valores dos pixels, de acordo com a Equação 2.1:

$$CDF(x) = \frac{1}{N \cdot M} \sum_{i=0}^{x} h(i), \tag{2.1}$$

onde, $CDF(x)$ é a função de distribuição acumulativa até o valor de pixel x, N e M são as dimensões da região, $h(i)$ é o valor do histograma no *bin* i.

Para evitar a amplificação de ruídos, o CLAHE limita o contraste redistribuindo o excesso de forma controlada, conforme descrito pela Equação 2.2:

$$h'(i) = \min\left(h(i), ClipLimit\right) + \frac{\sum_{i=0}^{L-1} \max\left(0, h(i) - ClipLimit\right)}{L}, \tag{2.2}$$

onde, $h'(i)$ é o histograma após a aplicação do limite de corte, $ClipLimit$ é o valor máximo permitido para cada *bin*, L é o número total de *bins* no histograma.

Após a equalização, a função de distribuição acumulativa é recalculada para remapear os valores dos pixels na imagem original. A Figura 3 ilustra o processo de aplicação do CLAHE.

A aplicação do CLAHE é particularmente relevante em radiografias da região lombo-sacra e dos pododáctilos, onde o baixo contraste pode dificultar a detecção de anomalias nas estruturas ósseas e articulares. Ao aprimorar o contraste, o CLAHE facilita a visualização de detalhes anatômicos críticos, melhorando a precisão diagnóstica e auxiliando os especialistas na identificação de patologias.

[4] Intervalo de intensidades de pixel.

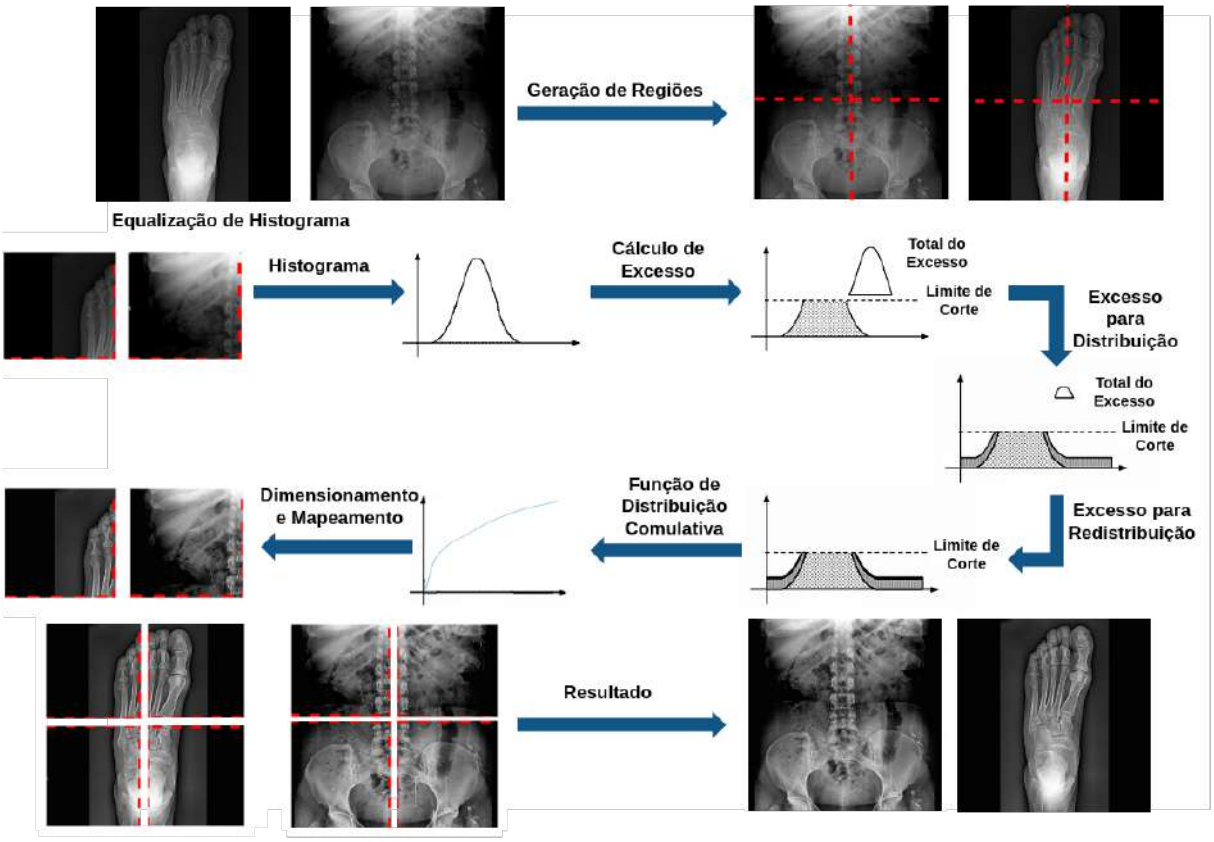


Figura 3 – Processo de aplicação do CLAHE em radiografias de lombo-sacra e pododáctilos. As imagens de entrada são divididas em regiões, e o limite de corte predefinido é aplicado para realizar a equalização do histograma.

Fonte: o próprio autor.

2.5 Aprendizado de Máquina

O aprendizado de máquina, conhecido como *machine learning*, é um subcampo essencial da engenharia e da ciência da computação, originado do estudo do reconhecimento de padrões e da teoria do aprendizado computacional em IA. Samuel (1959) definiu *machine learning* como o campo de estudo que dá aos computadores a habilidade de aprender sem serem explicitamente programados (SIMON, 2013). Essa definição captura a essência do *machine learning*, que é a capacidade de algoritmos aprenderem a partir de dados, ajustando-se e aprimorando seu desempenho sem intervenção humana direta (KOHAVI, 1998; BISHOP, 2006).

O *machine learning* envolve a construção de algoritmos que podem aprender com erros e fazer previsões sobre novos dados. Esses algoritmos constroem modelos a partir de exemplos de entrada, tomando decisões baseadas em dados, em vez de seguir regras programadas de forma rígida. Diferentemente da IA tradicional, que pode utilizar raciocínios indutivo e dedutivo, o *machine learning* foca principalmente no raciocínio indutivo, extraindo padrões e regras a partir de grandes conjuntos de dados (BISHOP,

2006).

O *machine learning* tem uma relação estreita com a estatística computacional, que se concentra em fazer previsões com o uso de computadores, além de compartilhar fortes conexões com a otimização matemática, fornecendo métodos e aplicações que complementam esse campo (WERNICK et al., 2010; ROUX; BENGIO; FITZGIBBON, 2011).

Esse campo é amplamente empregado em tarefas computacionais onde seria inviável criar algoritmos explícitos. Exemplos incluem filtragem de spam (SAAD; DARWISH; FARAJ, 2012), Reconhecimento Óptico de Caracteres (OCR) (RANJAN; BEHERA; REZA, 2021), processamento de linguagem natural (COLLOBERT et al., 2011), motores de busca (BRIN; PAGE, 1998), diagnósticos médicos(VIEIRA et al., 2021), bioinformática(GOLLERY, 2005), reconhecimento de fala e escrita(KARPAGAVALLI; CHANDRA, 2016), visão computacional (KRIZHEVSKY; SUTSKEVER; HINTON, 2017) e locomoção de robôs (KASHIRI et al., 2018).

Entre os algoritmos tradicionais de *machine learning*, destacam-se o *Support Vector Machines* (SVM), um classificador poderoso que busca o hiperplano ótimo que maximiza a margem de separação entre classes. O SVM é eficaz em problemas de alta dimensionalidade e é amplamente utilizado devido à sua robustez, mesmo quando há pouca separação entre classes (CORTES; VAPNIK, 1995).

Random Forest, um algoritmo de aprendizado baseado em múltiplas árvores de decisão. Ele combina as predições de várias árvores, cada uma treinada em subconjuntos aleatórios dos dados de treinamento, aumentando a acurácia e reduzindo o risco de *overfitting* (BREIMAN, 2001).

Multilayer Perceptron (MLP), uma classe de redes neurais artificiais composta por várias camadas de neurônios, permitindo ao MLP capturar e modelar relações complexas e não lineares entre as variáveis de entrada (POPESCU et al., 2009).

XGBoost, um algoritmo de *boosting* altamente eficiente, projetado para melhorar o desempenho em problemas de classificação e regressão. O XGBoost é amplamente utilizado por sua capacidade de lidar com dados desbalanceados e modelos complexos, além de oferecer alta precisão e velocidade (CHEN; GUESTRIN, 2016).

O *machine learning* é aplicado em várias tarefas, incluindo classificação, onde o objetivo é categorizar entradas em classes predefinidas (SARKER, 2021), regressão, em que as saídas são valores contínuos, como predição de preços ou temperaturas (MAULUD; ABDULAZEEZ, 2020), *clustering*, que agrupa dados não rotulados em *clusters* com base em semelhanças (SERRA-BURRIEL; AMES, 2022), e a redução de dimensionalidade, que simplifica dados de alta dimensionalidade para facilitar a visualização e análise (SERRA-BURRIEL; AMES, 2022).

Com o aumento da complexidade das tarefas de *machine learning*, técnicas mais avançadas, como o aprendizado profundo (*deep learning*), têm ganhado destaque. Na próxima seção, exploraremos o aprendizado profundo, que utiliza redes neurais profundas para modelar dados complexos e descobrir padrões intrincados em grandes volumes de dados.

2.6 Aprendizado Profundo

O aprendizado profundo (AP) é uma metodologia de *machine learning* que permite a construção de modelos computacionais capazes de aprender representações de dados em múltiplos níveis de abstração. Composta por diversas camadas de processamento, essa técnica tem sido amplamente utilizada em uma variedade de tarefas, como reconhecimento de fala (AL-JANABI; LATEEF, 2022), reconhecimento de texto (LONG; HE; YAO, 2021), reconhecimento visual de objetos (MANAKITSA et al., 2024), detecção de objetos (XIAO et al., 2020), simulação de buracos negros (DUARTE; NEMMEN; NAVARRO, 2022), descoberta de medicamentos (LI et al., 2023), e genômica (ALHARBI; RASHID, 2022). Modelos generativos baseados em arquiteturas como a de *Transformers*, incluindo a família GPT, também empregam aprendizado profundo para gerar textos, áudios e vídeos de forma autônoma (YENDURI et al., 2024). O AP permite a extração de estruturas complexas a partir de grandes volumes de dados, utilizando o algoritmo de retropropagação para ajustar os parâmetros internos do modelo, melhorando a representação em cada camada a partir das informações fornecidas pela camada anterior (LECUN YOSHUA BENGIO, 2015).

As Redes Neurais Artificiais (RNAs) são um exemplo clássico e eficaz de métodos de aprendizado profundo. Inspiradas nas redes neurais biológicas, as RNAs consistem em sistemas distribuídos e paralelos compostos por unidades de processamento simples, chamadas neurônios artificiais, que realizam cálculos matemáticos, geralmente não-lineares. Essas unidades estão organizadas em uma ou mais camadas e interconectadas por uma rede de conexões ponderadas, que armazenam o conhecimento do modelo e modulam a entrada recebida por cada neurônio (HAYKIN, 2007).

Um neurônio biológico é composto por dendritos, corpo celular e axônios. Os dendritos (entrada) recebem impulsos nervosos de outros neurônios e os conduzem ao corpo celular (processamento), que gera novos impulsos nervosos. Finalmente, os axônios (saída) transmitem esses impulsos para outros neurônios. A Figura 4 ilustra a estrutura de um neurônio biológico e suas funções principais (PURVES et al., 2010; MINSKY, 1988).

Analogamente, um neurônio artificial é composto por várias sinapses (conexões), cada uma associada a um peso. Um sinal de entrada X_j em uma sinapse j é multiplicado pelo peso correspondente W_k. Os sinais ponderados são somados, e essa soma é

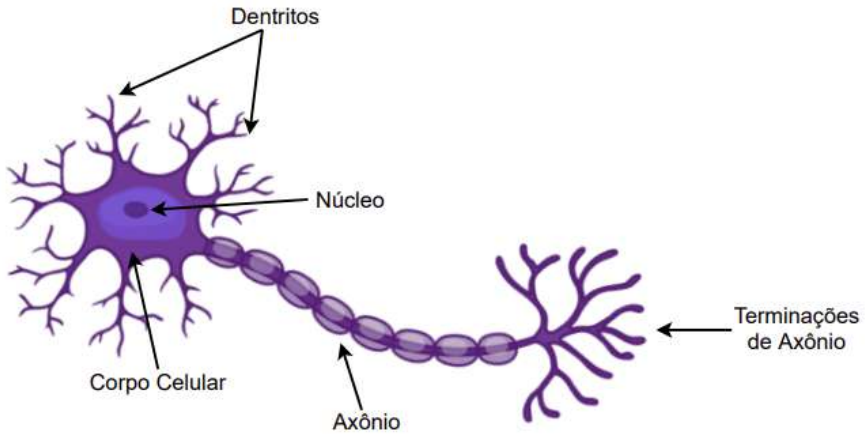


Figura 4 – Representação do neurônio biológico. O sinal percorre dos dendritos para os terminais dos axônios. O sinal é um pulso elétrico curto, chamado de potencial de ação ou pico.

Adptado de (VEXELS, 2017).

processada por uma função de ativação que restringe a amplitude da saída do neurônio. Uma entrada de viés (B_k) pode ajustar essa soma para aumentar ou diminuir a entrada da função de ativação (ROSENBLATT, 1958). A Figura 5 mostra a infraestrutura de um neurônio artificial, comparando suas partes com as de um neurônio biológico.

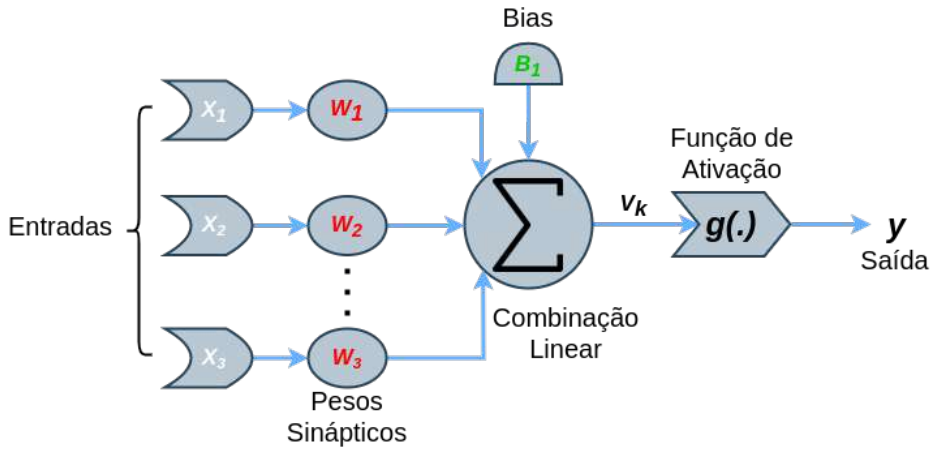


Figura 5 – Representação do neurônio artificial.

Inspirado em (VINICIUS, 2017).

Matematicamente, o processamento realizado por um neurônio artificial pode ser descrito pelas Equações 2.4 e 2.4.

$$V_k = \sum_{i=1}^{n} W_i * X_i + B_i, \tag{2.3}$$

$$y = g(V_k), \tag{2.4}$$

onde X_i são as entradas, W_i os pesos sinápticos, B_i o viés de ativação, $\sum_{i=1}^{n}$ a combinação linear, V_k o potencial de ativação, $g(\cdot)$ a função de ativação e y a saída.

2.7 Redes Neurais Convolucionais

As CNNs são uma classe sofisticadas de redes neurais profundas, amplamente aplicadas na análise de imagens. Essas redes, também conhecidas como redes neurais artificiais invariantes a deslocamentos ou invariantes espaciais, destacam-se por utilizar pesos compartilhados, o que promove a invariância translacional nas imagens processadas (ZHANG et al., 1990).

As CNNs são uma subcategoria das redes neurais multicamadas e compartilham o processo de treinamento supervisionado por meio da retropropagação (LECUN et al., 1998). A principal diferença em relação às redes neurais tradicionais está na sua arquitetura, que incorpora camadas de convolução responsáveis pelo pré-processamento das imagens de entrada. Essas camadas têm como objetivo extrair características relevantes, facilitando o processo de classificação.

A Figura 6 ilustra uma arquitetura genérica de uma CNN, destacando seus principais componentes como a camada de convolução, que realiza a extração de características; as camadas de *pooling*, que reduzem a dimensionalidade das características; e, finalmente, as camadas totalmente conectadas (CTC), ou Perceptron Multicamadas (PM), responsáveis pela classificação final.

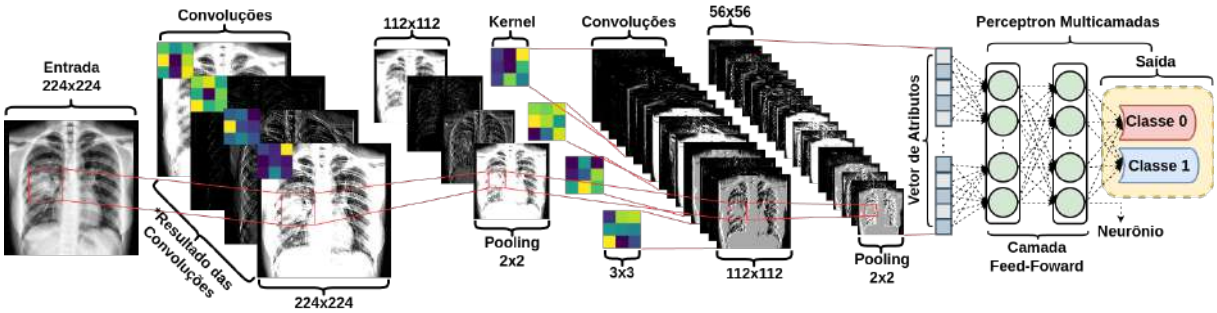


Figura 6 – Arquitetura CNN genérica baseada em Lecun et al. (1998). Aqui são ilustrados os principais processos de uma CNN e suas camadas.

Fonte: o próprio autor.

2.7.1 Redes Neurais Convolucionais Aplicadas

Neste trabalho, optou-se pela utilização de CNNs para a classificação de doenças nas radiografias de coluna lombo-sacra em imagens de radiografia, devido ao seu comprovado poder de generalização na análise de imagens médicas. Além disso, as CNNs foram empregadas na extração de características durante a etapa de geração

automática de laudos médicos, tanto para as imagens da coluna lombo-sacra quanto para as pododáctilos. Nosso levantamento do SOTA indicou que as CNNs são eficazes não apenas na classificação, mas também na extração de características de diversos tipos de imagens médicas.

Para este estudo, aplicamos quatro arquiteturas de CNNs distintas, sendo elas a Inception-V3 (IV3)(SZEGEDY et al., 2015), ResNet50 (RN50)(HE et al., 2015a), VGG16 (SIMONYAN; ZISSERMAN, 2014), e VGG19 (SIMONYAN; ZISSERMAN, 2015). Essas arquiteturas foram selecionadas com base em seu desempenho no desafio de banco de dados ImageNet (PARAS, 2017) e em suas características específicas, conforme detalhado na Tabela 1[5].

Tabela 1 – Características das arquiteturas de CNNs empregadas neste trabalho.

Modelo	Tamanho	Acc ImageNet	Tamanho de Entrada	Parâmetros	Profundidade	inferência (CPU)	inferência (GPU)
IV3	92 MB	0,937	(299×299)	23.851,784	159	42.2s	6.9s
RN50	98 MB	0,921	(224×224)	25.636,712	50	58.2s	4.6s
VGG16	528 MB	0,901	(224×224)	138.357,544	23	69.5s	4.2s
VGG19	549 MB	0,900	(224×224)	143.667.240	26	84.8s	4.4s

Embora todos esses modelos possuam milhões de parâmetros treináveis, capazes de capturar as características essenciais das imagens, a quantidade de parâmetros não é o único fator determinante para alcançar bons resultados. A forma como esses parâmetros são distribuídos nas camadas da rede, juntamente com as técnicas empregadas no treinamento, desempenha um papel crucial na eficácia e qualidade da inferência do modelo.

2.7.2 Camada de Convolução

Nas camadas de convolução, são utilizados filtros para processar as imagens e extrair características importantes, ao mesmo tempo que eliminam informações irrelevantes. Esses filtros, embora tenham a mesma profundidade que a imagem de entrada, são menores em termos de largura e altura. Por exemplo, para uma imagem 2D com dimensões $7 \times 7 \times 2$, o filtro pode ter dimensões $7 \times 7 \times 2$„ onde f pode ser 3, 5, 7, entre outros valores discretos. Na Figura 7, um *kernel* com $K = 3$ é ilustrado.

Durante a convolução, o filtro desliza sobre a imagem, realizando o produto escalar entre os pesos do filtro e os valores da imagem de entrada, gerando um mapa de ativação 2D. Esse mapa de ativação é uma imagem transformada que realça características específicas da entrada original, representada pela matriz resultante $I * K$[6]. Com isso, a CNN aprende características visuais importantes para tarefas como classificação,

[5] https://keras.io/api/applications/

[6] * é o operador de produto escalar.

segmentação ou detecção(SINGH; MEITEI; MAJUMDER, 2020; SINGH; MAJUMDER, 2020; GOODFELLOW YOSHUA BENGIO, 2016). A Figura 7 ilustra o processo de convolução.

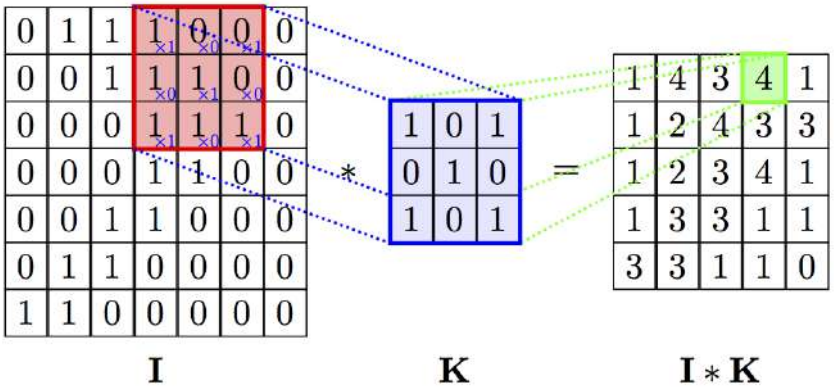


Figura 7 – Processo de convolução 2D. A matriz de entrada I, representando uma imagem, é processada pelo filtro de convolução K, resultando na matriz $I * K$.

Fonte: (LEE; GALLAGHER; TU, 2016).

Durante o processo de convolução 2D, a matriz de entrada I, que representa a imagem, é submetida ao filtro de convolução K. A operação gera uma nova matriz, resultante da aplicação do filtro, representada como $I * K$. Essa matriz transformada realça características relevantes da imagem original, facilitando a identificação de padrões importantes.

2.7.3 Camada de *Pooling*

Após a camada de convolução, a camada de *pooling* desempenha um papel importante na redução da dimensionalidade dos mapas de ativação, propagando eficientemente as características extraídas para as camadas subsequentes. Essa operação de subamostragem é essencial para mitigar o risco de *overfitting*, que ocorre quando um modelo se ajusta excessivamente aos dados de treinamento, perdendo a capacidade de generalizar para novos dados (MEANING. . . , 1933). Além disso, o *pooling* reduz significativamente a complexidade computacional ao diminuir o número de parâmetros que o modelo precisa aprender, aumentando a invariância às transformações de entrada (LEE; GALLAGHER; TU, 2016; SINGH; MAJUMDER, 2020; VIEIRA et al., 2021).

O processo de *pooling* envolve a aplicação de uma janela deslizante sobre os mapas de características, agregando informações de maneira a produzir uma representação mais compacta. As duas técnicas de *pooling* mais comuns são o *pooling* máximo e o *pooling* médio. No *pooling* máximo, o valor máximo dentro de cada janela é selecionado, capturando as características mais fortes da imagem. Já no *pooling* médio, a média dos valores dentro da janela é calculada, proporcionando uma

representação suavizada das características (SINGH; MEITEI; MAJUMDER, 2020; SCHERER ANDREAS MULLER, 2010).

A Figura 8 ilustra o processo de *pooling*, mostrando como janelas de 2x2 pixels são utilizadas para gerar mapas de características reduzidos. A figura também destaca as diferenças entre o *pooling* máximo e o *pooling* médio, demonstrando como cada método transforma a matriz de entrada em uma representação mais compacta.

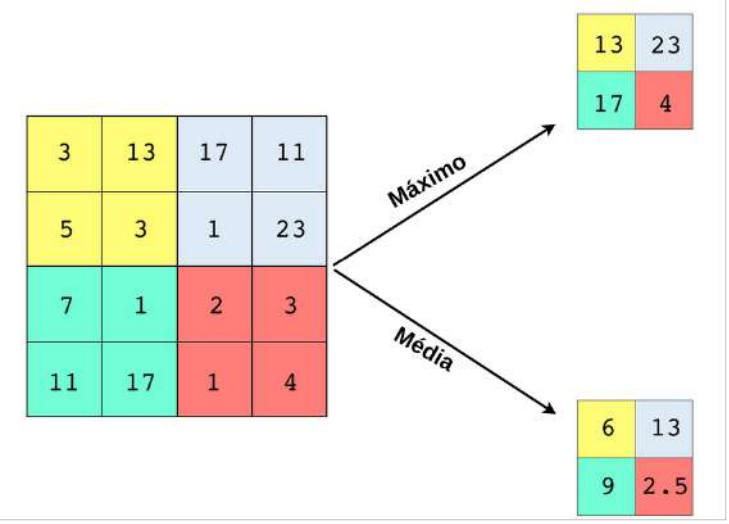


Figura 8 – Representação do processo de *pooling*. A matriz de entrada é processada utilizando *pooling* máximo e *pooling* médio com janelas de 2x2, resultando em matrizes reduzidas que preservam as características essenciais da imagem original.

Fonte: o próprio autor.

2.8 Redes Neurais Artificiais

As RNAs são sistemas computacionais inspirados na estrutura e funcionamento do cérebro humano, projetadas para modelar relações complexas entre entradas e saídas por meio de um processo de aprendizagem (BISHOP, 2006). Após a extração de características nas camadas de convolução e *pooling* em uma CNN, o resultado é geralmente alimentado em camadas totalmente conectadas, que são uma forma de RNA essencial para a etapa final de classificação.

As RNAs consistem em um conjunto de neurônios artificiais interconectados e organizados em camadas, que transformam entradas em saídas, gerando predições. Cada conexão entre neurônios possui um peso, ajustável durante o treinamento, permitindo que a rede aprenda padrões e características relevantes dos dados de entrada. Diferente de técnicas estatísticas tradicionais, as RNAs não fazem suposições prévias sobre a distribuição dos dados, o que as torna capazes de modelar funções não lineares e generalizar a partir de novos dados (GARDNER; DORLING, 1998).

Uma configuração comum de RNAs é a MLP, composta por pelo menos três camadas de neurônios. Uma camada de entrada, uma ou mais camadas ocultas e uma camada de saída. As MLPs são uma evolução dos Perceptrons, que adicionam camadas ocultas, permitindo resolver problemas não linearmente separáveis (ROSENBLATT, 1958; MINSKY, 1988). A camada de entrada recebe os dados, as camadas ocultas processam esses dados por meio de transformações não lineares, e a camada de saída gera as predições (GARDNER; DORLING, 1998). Cada neurônio nas camadas ocultas e de saída aplica uma função de ativação não linear, como a ReLU (*Rectified Linear Unit*)(AGARAP, 2018) ou a *Softmax*(GU et al., 2024), permitindo à rede capturar relações complexas nos dados (RUMELHART; HINTON; WILLIAMS, 1986).

A Figura 9 ilustra a estrutura de uma RNA, destacando as principais componentes, incluindo os pesos W_n, neurônios $\sum$, *bias* B_n e a função de ativação $g(\cdot)$. O processo de treinamento destas redes utiliza a técnica de retropropagação, onde o erro da predição é propagado de volta através da rede para ajustar os pesos, minimizando o erro ao longo do tempo através de métodos como a descida do gradiente (RUMELHART; HINTON; WILLIAMS, 1986).

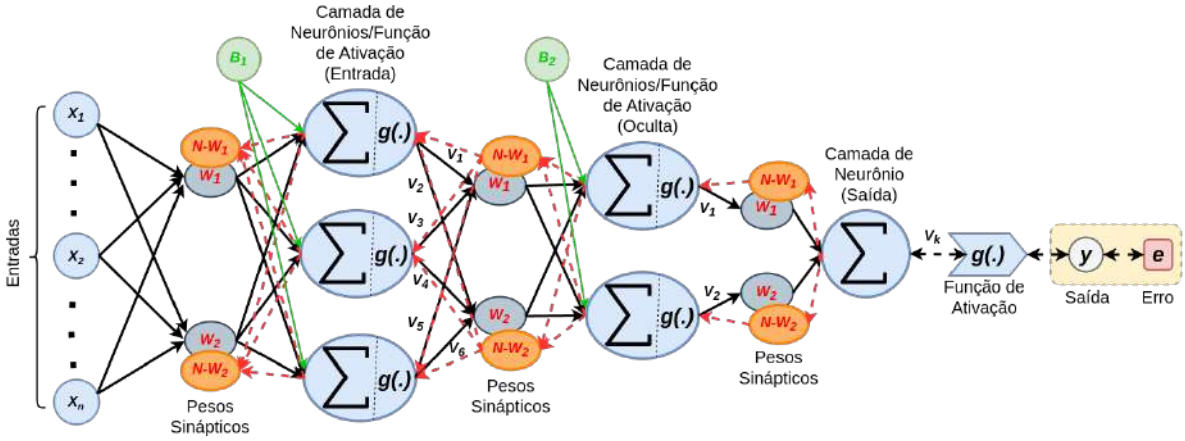


Figura 9 – Representação da RNA e suas principais partes, como os pesos W_n, neurônios $\sum$, *bias* B_n e a função de ativação $g(\cdot)$.

Fonte: o próprio autor.

As setas tracejadas em vermelho na Figura 9 representam a retropropagação dos erros, um processo crítico onde o erro e calculado na camada de saída é propagado de volta através da rede para ajustar os pesos $N - W_n$ de maneira a minimizar o erro de predição. Este processo iterativo continua até que a rede atinja um nível aceitável de precisão na tarefa de classificação ou regressão. Apesar da eficácia das RNAs na modelagem de problemas complexos, a adição de mais camadas e neurônios aumenta significativamente o custo computacional e a complexidade do modelo (RUMELHART; HINTON; WILLIAMS, 1986).

2.9 Transferência de Aprendizado e Ajuste Fino

A Transferência de Aprendizado (TA) é uma técnica essencial em AP, onde um modelo pré-treinado em um conjunto de dados extenso é reutilizado para uma nova tarefa em um domínio diferente. Essa abordagem é amplamente empregada em tarefas como classificação, detecção de objetos e segmentação de imagens, utilizando CNNs para melhorar o desempenho em problemas específicos a partir de modelos previamente treinados em problemas mais gerais (PARAS, 2017; YOSINSKI et al., 2014; LI; HOIEM, 2016; RAZAVIAN et al., 2014). Um exemplo comum é o uso de redes pré-treinadas em conjuntos de dados como o ImageNet (RUSSAKOVSKY et al., 2015), que demonstram desempenho superior em tarefas de classificação de imagens médicas.

Após a TA, realiza-se o Ajuste Fino (*fine-tuning*), onde os parâmetros do modelo pré-treinado são ajustados para uma tarefa específica. Nesse processo, as camadas totalmente conectadas da rede original são geralmente substituídas por novas camadas, que são treinadas com os novos dados. Existem duas abordagens principais o ajuste fino raso, onde apenas as novas camadas são treinadas, mantendo os pesos das camadas convolucionais inalterados, e o ajuste fino profundo, onde toda a rede é treinada, permitindo que as camadas convolucionais se adaptem às novas características do conjunto de dados (JOHN., 1998; YOSINSKI et al., 2014; RAZAVIAN et al., 2014).

Estudos mostram que o ajuste fino de CNNs pré-treinadas oferece uma alternativa prática ao treinamento do zero, permitindo alcançar um desempenho superior com menos dados e menor tempo de treinamento (TAJBAKHSH et al., 2016; VIEIRA et al., 2021). O ajuste fino não se limita a CNNs, mas pode ser aplicado a outras arquiteturas, como *Transformers*, demonstrando sua flexibilidade e eficácia em uma ampla gama de aplicações. A Figura 10 ilustra os processos de ajuste fino raso e profundo.

2.10 *Transformers*

Transformers são modelos de linguagem baseados em redes neurais que têm mostrado resultados notáveis em tarefas de Processamento de Linguagem Natural (*Natural Language Processing (NLP)*), destacando-se por sua capacidade de compreender relações contextuais complexas em sequências, como palavras e frases (VASWANI et al., 2017; BROWN et al., 2020). Na desafiadora tarefa de gerar legendas para imagens, conhecida como *image-to-text*, os *Transformers* são amplamente adotados devido à sua eficácia em capturar dependências de longo alcance nos dados (PAVLOPOULOS et al., 2022).

O treinamento de modelos de *Transformers* requer uma etapa de pré-processamento dos dados. Em tarefas de *image-to-text*, por exemplo, as imagens são convertidas em representações numéricas podendo utilizar CNNs para extração de representações

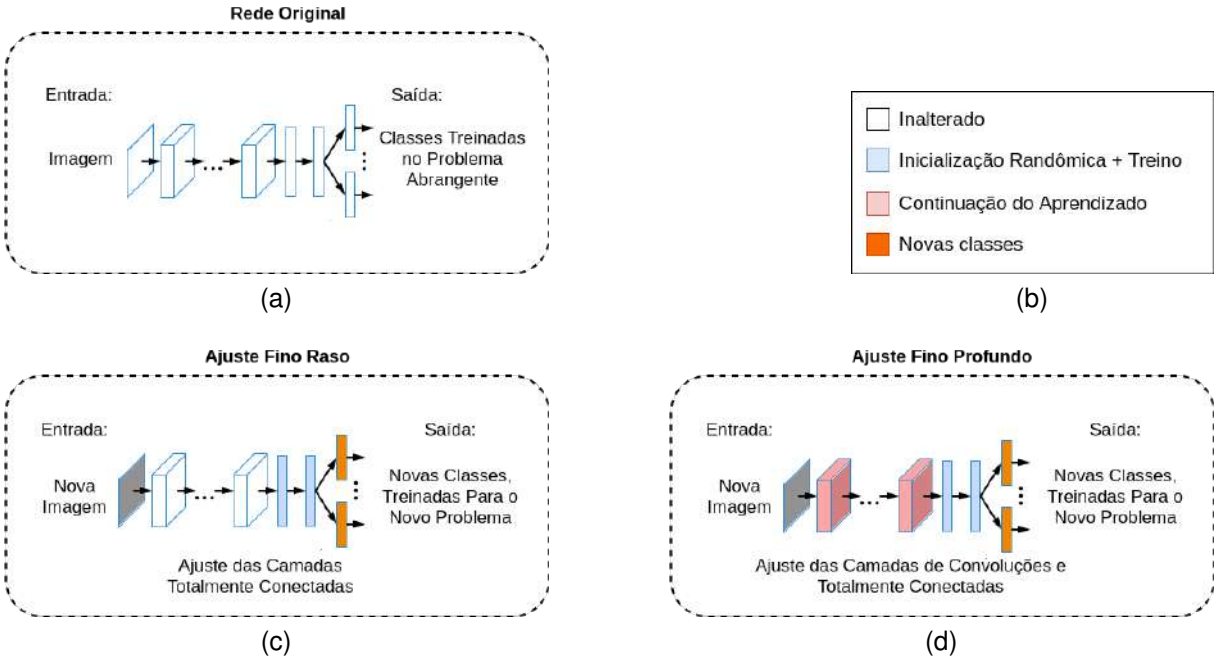


Figura 10 – Ilustração do processo de Transferência de Aprendizado para o Ajuste Fino: (a) Uma arquitetura original treinada em um conjunto de dados abrangente; (c) Transferência de Aprendizado para o processo de Ajuste Fino Raso; (d) Transferência de Aprendizado para o processo de Ajuste Fino Profundo.

Adaptado de (LI; HOIEM, 2016).

vetoriais, enquanto as legendas são tokenizadas e codificadas em vetores numéricos (VASWANI et al., 2017). Esse processo permite que o *transformer* aprenda a mapear representações visuais complexas para sequências de texto coerentes.

Após o treinamento, o modelo pode ser utilizado para gerar legendas para novas imagens não vistas anteriormente. Durante a inferência, a imagem é processada pela mesma CNN usada no treinamento para obter sua representação vetorial. Essa representação é então combinada com a saída do modelo de *transformer*, utilizando *embeddings* e uma camada *Softmax* para produzir a legenda final (VASWANI et al., 2017; PAVLOPOULOS et al., 2022). A Figura 11 ilustra a arquitetura do *transformer*.

Recentemente, modelos codificador-decodificador que utilizam *Transformers* tanto no codificador quanto no decodificador têm sido aplicados na geração de laudos médicos preliminares. O codificador *Transformer* processa características extraídas de um codificador de imagens baseado em CNN. O decodificador *Transformer* é aprimorado com um mecanismo de memória relacional, que ajuda a “lembrar” padrões de texto recorrentes (por exemplo, “os pulmões estão claros”) que aparecem em relatórios de imagens similares. Este mecanismo, inspirado em portas de entrada e esquecimento semelhantes às células LSTM, melhora a capacidade do modelo de gerar relatórios precisos e contextualmente relevantes. Este método pioneiro no uso de *Transformers* para relatórios médicos alcançou resultados promissores (PAVLOPOULOS et al., 2022).

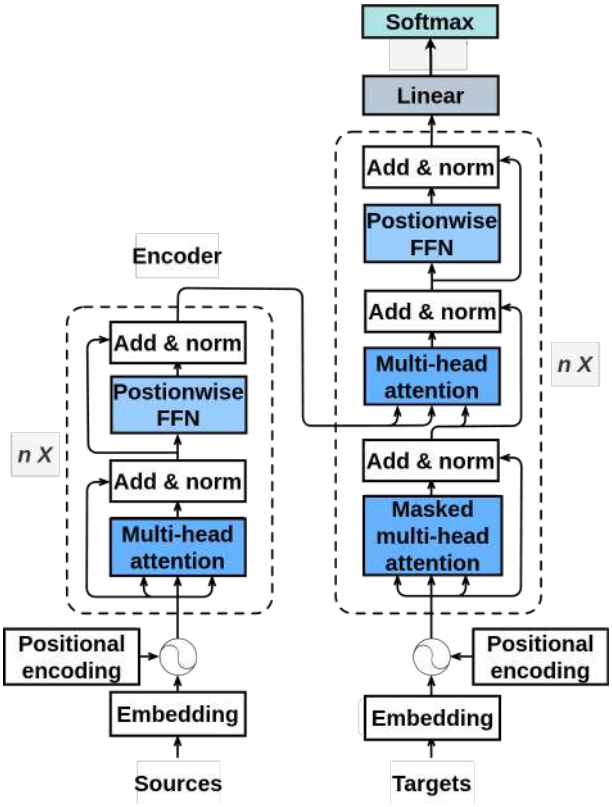


Figura 11 – Estrutura da arquitetura *Transformer*, destacando seus componentes principais como o codificador (*encoder*) e o decodificador (*decoder*). No contexto de modelos *image-to-text*, as imagens são transformadas em vetores numéricos, enquanto as legendas são convertidas em vetores deslocados. O codificador processa as representações numéricas das imagens, gerando representações contextuais que são então utilizadas pelo decodificador para gerar as sequências de saída correspondentes às legendas.

Adaptado de (VASWANI et al., 2017).

2.10.1 Codificador do *Transformers*

O codificador da arquitetura *Transformer* desempenha um papel fundamental no processamento de sequências de entrada para diversas tarefas de NLP. Ele é composto por uma série de camadas de autoatenção seguidas por redes neurais densas. Cada camada do codificador utiliza mecanismos de autoatenção multi-cabeça para capturar dependências entre diferentes palavras na sequência de entrada. As conexões residuais e a normalização por camada são empregadas para estabilizar o processo de treinamento e melhorar a convergência (VASWANI et al., 2017).Na Figura 11 podemos visualizar a arquitetura do codificador.

A camada de Autoatenção Multi-cabeça permite que o modelo foque em diferentes partes da sequência de entrada simultaneamente, proporcionando uma compreensão mais rica e contextualizada dos dados. Este processo pode ser descrito pelas seguintes

operações:

Primeiro, são calculados os vetores de consulta (Query), chave (Key) e valor (Value) como ilustra a Equação 2.5:

$$\text{Query} = \boldsymbol{X} \cdot \boldsymbol{W}_Q, \quad \text{Key} = \boldsymbol{X} \cdot \boldsymbol{W}_K, \quad \text{Value} = \boldsymbol{X} \cdot \boldsymbol{W}_V \tag{2.5}$$

Em seguida, os vetores de consulta, chave e valor são divididos em múltiplas cabeças. Para cada cabeça, o escore é calculado da seguinte forma na Equação 2.6:

$$\text{Score}_i = \text{softmax}\left(\frac{\text{Query}_i \cdot \text{Key}_i^T}{\sqrt{d_{\text{head}}}}\right) \cdot \text{Value}_i \quad \text{para } i = 1 \text{ a } N_{\text{heads}} \tag{2.6}$$

As saídas dos escores de todas as cabeças são então concatenadas.

Após a camada de Autoatenção, aplica-se uma Conexão Residual e uma Normalização por camada para manter a estabilidade do treinamento, apontado na Equação2.7:

$$\text{Output} = \text{LayerNorm}(\boldsymbol{X} + \text{Dropout}(\text{Concatenated Score})) \tag{2.7}$$

Cada camada do codificador inclui uma Rede Neural Feedforward, que aplica uma transformação não linear aos dados processados na Equação 2.8:

$$\text{OutputFFN} = \text{ReLU}(\text{Output} \cdot \boldsymbol{W}_1 + \boldsymbol{b}_1) \cdot \boldsymbol{W}_2 + \boldsymbol{b}_2 \tag{2.8}$$

Mais uma vez, uma Conexão Residual e uma Normalização por camada são aplicadas para a saída da Rede Feedforward, como mostra a Equação2.9:

$$\text{Output} = \text{LayerNorm}(\text{Output} + \text{Dropout}(\text{OutputFFN})) \tag{2.9}$$

A Saída Final do Codificador é então obtida pela aplicação da Normalização por camada na soma Residual da entrada e a atenção computada na Equação 2.10:

$$\boldsymbol{X}_{\text{encoder}} = \text{LayerNorm}(\boldsymbol{X} + \text{Dropout}(\text{Attention}(\boldsymbol{X}))) \tag{2.10}$$

As operações acima descrevem os principais componentes do codificador do *Transformer*, destacando seu mecanismo para extrair representações contextualizadas das sequências de entrada. A Equação 2.10 representa a saída da camada do codificador. Em nosso trabalho, optamos por utilizar um codificador com uma arquitetura composta

por 12 camadas de autoatenção multi-cabeça.

Transformers têm servido como a base para muitos modelos de última geração em NLP (WOLF et al., 2020), como BERT (Bidirectional Encoder Representations from *Transformers*) (DEVLIN et al., 2018), GPT (*Generative Pre-trained Transformer*) (YENDURI et al., 2024) e RoBERTa (LIU et al., 2019). Estes modelos têm alcançado resultados de ponta em diversas tarefas de NLP, incluindo tradução automática, resumo de texto, geração de texto e resposta a perguntas dentre outras atividades.

2.10.2 Decodificador do *Transformers*

O decodificador do *Transformer* é responsável por gerar sequências de saída a partir das representações contextualizadas produzidas pelo codificador. Ele consiste em uma série de camadas de autoatenção, tanto para a sequência de entrada (saída do codificador) quanto para as posições anteriores na sequência de saída. Semelhante ao codificador, o decodificador utiliza conexões residuais e normalização de camada para garantir um treinamento estável e eficaz (VASWANI et al., 2017). Na Figura 11 podemos visualizar a arquitetura do decodificador.

A camada de Autoatenção no Decodificador permite que o modelo foque em diferentes partes da sequência de saída já gerada, garantindo que a geração de cada palavra considere as palavras precedentes de forma eficiente. Este processo é descrito pelas seguintes operações:

Primeiro, os vetores de consulta (Query), chave (Key) e valor (Value) são calculados para a entrada do decodificador como mostra as Equações 2.11, 2.12 e 2.13:

$$\text{Query} = \text{DecoderInput} \cdot \boldsymbol{W}_{Q1}, \tag{2.11}$$

$$\text{Key} = \text{DecoderInput} \cdot \boldsymbol{W}_{K1}, \tag{2.12}$$

$$\text{Value} = \text{DecoderInput} \cdot \boldsymbol{W}_{V1} \tag{2.13}$$

Em seguida, o escore é calculado com a Equação 2.14:

$$\text{Score} = \text{softmax}\left(\frac{\text{Query} \cdot \text{Key}^T}{\sqrt{d_{\text{head}}}}\right) \cdot \text{Value} \tag{2.14}$$

Nesta camada, a Atenção é aplicada entre a entrada do Decodificador e a saída do Codificador, permitindo que o Decodificador acesse informações contextuais da sequência de entrada pelas Equações 2.15, 2.16 e 2.17:

$$\text{Query} = \text{DecoderInput} \cdot \boldsymbol{W}_{Q2}, \tag{2.15}$$

$$\text{Key} = \text{EncoderOutput} \cdot \boldsymbol{W}_{K2}, \tag{2.16}$$

$$\text{Value} = \text{EncoderOutput} \cdot \boldsymbol{W}_{V2} \tag{2.17}$$

O escore é novamente calculado da seguinte maneira pela Equação 2.18:

$$\text{Score} = \text{softmax}\left(\frac{\text{Query} \cdot \text{Key}^T}{\sqrt{d_{\text{head}}}}\right) \cdot \text{Value} \tag{2.18}$$

A camada de Autoatenção Multi-cabeça permite ao modelo focar em diferentes partes da sequência de entrada e saída simultaneamente, podemos ver o processo na Equação 2.19:

$$\text{Score}_i = \text{softmax}\left(\frac{\text{Query}_i \cdot \text{Key}_i^T}{\sqrt{d_{\text{head}}}}\right) \cdot \text{Value}_i \quad \text{para } i = 1 \text{ a } N_{\text{heads}} \tag{2.19}$$

As saídas dos escores de todas as cabeças são então concatenadas. Similar ao codificador, a saída das camadas de Atenção no Decodificador é estabilizada através de Conexões Residuais e Normalização por camada. A Equação 2.20 ilustra o processo:

$$\text{Output} = \text{LayerNorm}(\text{DecoderInput} + \text{Dropout}(\text{Concatenated Score})) \tag{2.20}$$

Cada camada do Decodificador inclui uma Rede Neural *Feedforward* para aplicar transformações não lineares adicionais pela Equação 2.21:

$$\text{FFNOutput} = \text{ReLU}((\text{Output} \cdot \boldsymbol{W}_1) + \boldsymbol{b}_1) \cdot \boldsymbol{W}_2 + \boldsymbol{b}_2 \tag{2.21}$$

Mais uma vez, uma Conexão Residual e Normalização por camada são aplicadas para a Saída da Rede Feedforward atráves da Equação 2.22:

$$\text{Output} = \text{LayerNorm}(\text{Output} + \text{Dropout}(\text{FFNOutput})) \tag{2.22}$$

A Saída Final do Decodificador é então obtida pela aplicação da Normalização

por camada na Soma Residual da entrada e a Atenção computada, seguida de uma camada *Softmax* para gerar as probabilidades das próximas palavras na sequência pela Equação 2.22:

$$\text{DecoderOutput} = \text{Softmax}(\text{Output}) \tag{2.23}$$

A Saída Final do Decodificador é então obtida pela aplicação da Normalização por camada na Soma Residual da entrada e a Atenção computada, seguida de uma camada *Softmax* para gerar as probabilidades das próximas palavras na sequência pela Equação 2.24:

$$\text{DecoderOutput} = \text{Softmax}(\text{Output}) \tag{2.24}$$

As operações acima, numeradas de 2.11 a 2.24, descrevem os principais componentes do decodificador do *Transformer*, ilustrando seu mecanismo para gerar sequências de saída baseadas nas representações contextualizadas do codificador. A Equação 2.24 representa a saída da camada do decodificador.

2.11 Aumento de Dados

Uma das maiores dificuldades na classificação, segmentação e detecção de imagens médicas é a escassez de amostras. Este desafio pode ser particularmente crítico em exames de lombo-sacra e pododáctilos, onde a variabilidade anatômica e a presença de patologias específicas exigem um volume considerável de dados para um treinamento robusto de modelos de IA generativa, especialmente aqueles baseados em arquiteturas de *Transformers*.

A falta de diversidade nos dados podem prejudicar o aprendizado de modelos de CNNs e *Transformers*. Uma estratégia eficaz para mitigar esse problema é o Aumento de Dados (AD), que envolve a criação de amostras sintéticas a partir de amostras reais. O AD compreende um conjunto de técnicas que aumentam tanto o tamanho quanto a qualidade dos conjuntos de dados usados no treinamento de modelos, aprimorando a capacidade de generalização dos modelos de IA (CONNOR, 2019).

Diversas técnicas podem ser empregadas para o AD em imagens médicas, incluindo transformações geométricas, ajustes no espaço de cores, aplicação de filtros de convolução (GONZALEZ, 2011), mistura de imagens (DABOUEI et al., 2020), apagamento aleatório (ZHONG et al., 2017), aumento do espaço de características (LIU et al., 2018), uso de redes adversárias generativas (*Generatives Artificiais Networks* GANs)(TRAN et al., 2021; YI; WALIA; BABYN, 2019), transferência de estilo(ZHENG et al., 2019), e esquemas de meta-aprendizagem (ANTONIOU; STORKEY, 2019). Essas técnicas

visam melhorar a invariância dos modelos, um dos maiores desafios nas tarefas de reconhecimento de padrões em imagens (CONNOR, 2019).

Embora o aumento de dados possa melhorar significativamente a precisão na reconhecimento de padrões em imagens, sua aplicação é complexa. A implementação manual dessas técnicas é trabalhosa, e as possibilidades para a geração de novas imagens sintéticas são vastas. Além disso, cada técnica de AD possui parâmetros específicos, como a rotação, que requer a definição do grau e do sentido da rotação (CUBUK et al., 2018; CUBUK et al., 2019).

A adoção em larga escala dessas técnicas enfrenta desafios na definição de quais métodos empregar, dado o amplo espaço de busca dos parâmetros. Uma estratégia eficiente envolve a identificação das métricas mais adequadas ao problema e a determinação dos espaços de busca aceitáveis para os parâmetros, evitando distorções indesejadas nas imagens. Isso exige um conhecimento profundo do problema a ser resolvido e um estudo detalhado das imagens resultantes do processo de aumento para uma tomada de decisão mais assertiva (VIEIRA et al., 2021).

Ao aplicar as funções de AD, combinando-as e ajustando os parâmetros de forma randômica, é possível gerar um número significativamente maior de amostras sintéticas. Cada imagem original pode gerar várias outras, limitadas apenas pelo número de funções de aumento e pelo espaço de busca dos parâmetros dessas funções (CUBUK et al., 2019).

Além disso, os modelos de *Transformers*, que têm mostrado grande eficácia em diversas aplicações de IA, dependem de grandes quantidades de dados para atingir seu pleno potencial. Estes modelos, devido à sua capacidade de capturar dependências de longo alcance e compreender complexas relações entre os dados, são particularmente adequados para tarefas que exigem um grande volume de dados, como aquelas encontradas em imagens médicas de lombo-sacra e pododáctilos. A necessidade de grandes conjuntos de dados torna a prática de aumento de dados uma abordagem crucial, pois ela permite expandir artificialmente o volume de amostras disponíveis, sem a necessidade de gerar novas amostras do zero, o que pode ser dispendioso e demorado (SHAMSHAD et al., 2023).

2.12 Métricas de Validação

A avaliação de modelos é uma etapa importante no desenvolvimento de metodologias eficazes de *machine learning*, pois permite verificar o desempenho de generalização do modelo em relação a dados não vistos durante o treinamento.

Para a pesquisa relacionada à classificação de imagens da coluna lombo-sacra, optamos por empregar um conjunto de teste fixo e independente, utilizado para medir o

desempenho de generalização do modelo. Essa escolha não apenas permite uma avaliação precisa da capacidade do modelo em lidar com casos futuros desconhecidos, mas também se deu em função das limitações de tempo e dos recursos computacionais disponíveis durante o período do estudo. Essa abordagem foi preferida em detrimento de técnicas como reamostragem ou validação de espera, devido ao seu custo computacional elevado e à necessidade de garantir uma previsão eficiente da aplicabilidade do modelo a novos dados.

Na avaliação dos métodos empregados para a geração automatizada de laudos previos, utilizamos a técnica de validação cruzada *k-fold*(KOHA, 1995). Nessa técnica, o conjunto de dados é dividido aleatoriamente em *k* subconjuntos de tamanhos aproximadamente iguais. A cada iteração, um subconjunto é reservado para validação, enquanto os *k-1* subconjuntos restantes são utilizados para treinamento. O processo é repetido *k* vezes, garantindo que cada subconjunto seja usado exatamente uma vez como conjunto de validação. Nesta etapa, os recursos computacionais e o tempo disponíveis eram superiores a etapa citada anteriormente, o que permitiu utilizar a metodologia de validação cruzada *k-fold*, proporcionando uma avaliação mais robusta do desempenho do modelo generativo. A Figura 12 ilustra esse processo.

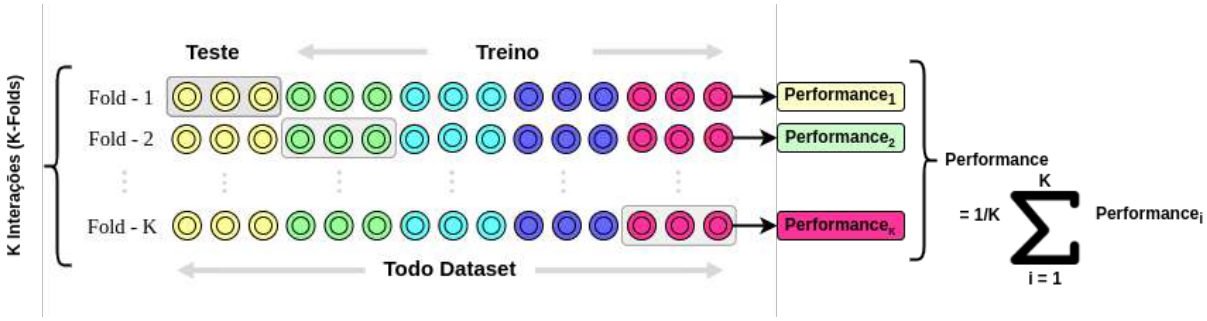


Figura 12 – Ilustração da validação cruzada em compartimentos.

Fonte: o próprio autor.

Os resultados de todas as iterações são agregados para fornecer uma avaliação completa do desempenho do modelo. A média dos desempenhos em todas as iterações é utilizada como índice de avaliação. Embora essa abordagem seja computacionalmente mais custosa do que a tradicional separação estática entre conjuntos de treinamento e teste, ela permite uma exploração mais detalhada do conjunto de dados e é especialmente vantajosa quando o número de amostras é limitado (SARAIVA1 et al., 2019). Essa técnica também oferece uma medida da capacidade do modelo de generalizar para dados desconhecidos.

Para a avaliação do desempenho, consideramos quatro situações possíveis:

O Verdadeiro Positivo (VP) ocorre quando uma imagem da classe ANORMAL é

corretamente classificada. O Falso Positivo (FP) acontece quando uma imagem de outra classe é incorretamente classificada como positiva para ANORMAL. O Verdadeiro Negativo (VN) refere-se a uma imagem de outra classe que é corretamente classificada como negativa para ANORMAL. Por fim, o Falso Negativo (FN) é quando uma imagem da classe ANORMAL é incorretamente classificada como pertencente a outra classe.

Empregamos as seguintes métricas clássicas de validação a acurácia, que reflete a proporção de predições corretas realizadas por um modelo (CONGALTON, 1986); o F-Score, que combina sensibilidade e precisão, oferecendo uma medida equilibrada entre essas duas dimensões (RIJSBERGEN, 1979); a curva ROC (Receiver Operating Characteristic), amplamente utilizada para avaliar o desempenho de modelos de classificação em diferentes limiares de decisão; a área sob a curva ROC (AUC), que fornece uma medida geral da capacidade discriminativa do modelo (HAND, 2001; CAIO et al., 2019); a precisão, que quantifica a taxa de acerto na identificação da classe positiva (OLSON, 2008); a sensibilidade, que avalia a capacidade do modelo de identificar corretamente as amostras positivas (THARWAT, 2020); a especificidade, que mede a habilidade do modelo em identificar corretamente as amostras negativas (THARWAT, 2020); e o índice *Kappa* (κ), que mede o grau de concordância entre proporções derivadas de amostras dependentes (COHEN, 1968; FLEISS; COHEN, 1973).

Além dessas métricas, utilizamos também a taxa de descoberta falsa (FDR) e a taxa de omissão falsa (FOR), conforme as Equações 2.25 e 2.26, respectivamente,

$$FDR = \frac{FP}{FP + TP}, \tag{2.25}$$

$$FOR = \frac{FN}{FN + VN}. \tag{2.26}$$

Para as tarefas de segmentação, utilizamos métricas como o Coeficiente de Similaridade de Dice e o Coeficiente de Similaridade de Jaccard (SHRIVASTAVA; BHARTI, 2016). Essas métricas permitem avaliar a sobreposição entre as regiões segmentadas pelo modelo e as regiões de referência, fornecendo uma medida quantitativa da precisão da segmentação. O Coeficiente de Similaridade de *Dice* é calculado conforme a Equação 2.27:

$$\text{Dice} = \frac{2|A \cap B|}{|A| + |B|}, \tag{2.27}$$

onde $|A|$ e $|B|$ são os tamanhos dos conjuntos A e B, respectivamente, e $|A \cap B|$ é o tamanho da interseção dos conjuntos A e B.

O Coeficiente de Similaridade de Jaccard é calculado conforme a Equação 2.28,

$$\text{Jaccard} = \frac{|A \cap B|}{|A \cup B|}, \tag{2.28}$$

onde $|A \cup B|$ é o tamanho da união dos conjuntos A e B.

Para a avaliação da IA generativa do tipo *image-to-text*, especialmente na geração de laudos médicos, utilizamos diversas métricas de validação de textos amplamente reconhecidas na literatura, como BLEU (PAPINENI et al., 2002), METEOR (DENKOWSKI; LAVIE, 2014) e ROUGE (LIN, 2004), cada uma oferecendo perspectivas complementares sobre a qualidade e a precisão dos textos gerados.

A métrica BLEU (*Bilingual Evaluation Understudy*) avalia a sobreposição de n-gramas entre as sequências de texto geradas e as referências fornecidas por especialistas, conforme a Equação 2.29 (PAPINENI et al., 2002),

$$BLEU = BP \times \exp\left(\sum_{n=1}^{N} w_n \log p_n\right), \tag{2.29}$$

onde BP é o *brevity penalty*, w_n são os pesos de n-gramas, e p_n são as precisões de n-gramas.

A métrica METEOR (*Metric for Evaluation of Translation with Explicit ORdering*) considera não apenas n-gramas, mas também sinônimos e reordenações, conforme a Equação 2.30 (DENKOWSKI; LAVIE, 2014),

$$METEOR = Fmean \times (1 - Penalty), \tag{2.30}$$

onde $Fmean$ é uma média ponderada de precisão e recall, e $Penalty$ é uma penalidade baseada em fragmentação.

A métrica ROUGE (*Recall-Oriented Understudy for Gisting Evaluation*) avalia a sobreposição de n-gramas, palavras ou sentenças entre o texto gerado e a referência, conforme a Equação 2.31 (LIN, 2004),

$$ROUGE - N = \frac{\sum_{S \in \{Referência\}} \sum_{grama_n \in S} Contagem\_correta(grama_n)}{\sum_{S \in \{Referência\}} \sum_{grama_n \in S} Contagem\_total(grama_n)}, \tag{2.31}$$

As métricas BLEU, METEOR e ROUGE-L são amplamente utilizadas para avaliar a qualidade de texto gerado por modelos de tradução automática, resumo de textos e sistemas de geração de legendas. Essas métricas comparam frases candidatas com uma frase de referência (original) para medir a similaridade entre elas. No exemplo apresentado na Tabela 2, a frase original foi comparada com cinco diferentes frases candidatas sendo uma idêntica, uma semanticamente próxima, duas com desvios mais

significativos e uma que apresenta descrições totalmente diferentes.

Tabela 2 – Exemplos de comparação das métricas BLEU (1-4), METEOR e ROUGE-L para diferentes frases candidatas.

Tipo	Frase	BLEU-1	BLEU-2	BLEU-3	BLEU-4	METEOR	ROUGE-L
Original	"MÍNIMO DESVIO DO EIXO LOMBAR PARA ESQUERDA, PROVAVELMENTE POSTURAL-POSICIONAL. CORPOS VERTEBRAIS DE FORMA E CONTORNOS ANATÔMICOS. PENDÍCULOS E APÓFISE TRANSVERSAS EM ALTERAÇÕES. DISCOS VERTEBRAIS CONSERVADOS."	-	-	-	-	-	-
Candidato (A)	"MÍNIMO DESVIO DO EIXO LOMBAR PARA ESQUERDA, PROVAVELMENTE POSTURAL-POSICIONAL. CORPOS VERTEBRAIS DE FORMA E CONTORNOS ANATÔMICOS. PENDÍCULOS E APÓFISE TRANSVERSAS EM ALTERAÇÕES. DISCOS VERTEBRAIS CONSERVADOS."	0.999	0.999	0.999	0.999	0.999	0.999
Candidato (B)	"LEVE DESVIO DO EIXO LOMBAR PARA A ESQUERDA, POSSIVELMENTE POSTURAL. CORPOS VERTEBRAIS DE FORMA E CONTORNOS ESTÃO ANATÔMICOS. PENDÍCULOS E APÓFISES TRANSVERSAIS SEM ALTERAÇÕES. DISCOS VERTEBRAIS ESTÃO CONSERVADOS."	0.678	0.571	0.712	0.483	0.407	0.724
Candidato (C)	"GRANDE DESVIO DO EIXO LOMBAR PARA A DIREITA, DE ORIGEM ESTRUTURAL. CORPOS VERTEBRAIS COM FORMAS ALTERADAS. PENDÍCULOS E APÓFISES TRANSVERSAS APRESENTAM DEFORMAÇÕES. DISCOS VERTEBRAIS COMPROMETIDOS."	0.436	0.301	0.158	0.000	0.233	0.420
Candidato (D)	"OSTEÓFITOS MARGINAIS E DESGASTE SIGNIFICATIVO DOS DISCOS INTERVERTEBRAIS EM L4-L5. EVIDÊNCIA DE DEGENERAÇÃO ARTICULAR E LEVE ESTENOSE DO CANAL VERTEBRAL."	0.233	0.000	0.000	0.000	0.095	0.087

A frase candidata (A), que é idêntica à original, alcançou a pontuação máxima em todas as métricas (BLEU-1, BLEU-2, BLEU-3, BLEU-4, METEOR e ROUGE-L), confirmando sua elevada similaridade com a frase de referência. A frase candidata (B), que é semanticamente próxima, mas contém algumas alterações na estrutura e na escolha das palavras, obteve pontuações ligeiramente inferiores em todas as métricas, refletindo uma leve perda de similaridade.

As frases candidatas (C) e (D), que apresentam desvios semânticos mais significativos, especialmente no sentido do desvio do eixo lombar, registraram uma queda substancial em todas as métricas. BLEU-4 e ROUGE-L foram particularmente sensíveis a essas alterações, evidenciando as diferenças estruturais e de sequência nas palavras.

Por fim, a quinta frase candidata, que descreve condições completamente

diferentes, obteve as piores pontuações em todas as métricas. Isso demonstra claramente a eficácia dessas métricas em detectar grandes variações na qualidade e similaridade entre o texto gerado e a frase original, com BLEU focando na correspondência de n-gramas, METEOR na correspondência semântica e ROUGE-L na similaridade de estrutura e ordenação das palavras.

2.13 Considerações Finais

Neste capítulo, foi apresentada a fundamentação teórica necessária para a compreensão das técnicas utilizadas e suas aplicações no método proposto. Foram abordados temas como conceitos básicos de classificação de imagens e geração de laudos médicos, técnicas de aprendizado profundo, métricas de validação de modelos, e os procedimentos de pré-processamento digital de imagens. Esta base teórica é essencial para o desenvolvimento e a avaliação dos modelos de IA aplicados ao problema em questão.

3 Trabalhos Relacionados

Este capítulo apresenta os principais trabalhos relacionados ao desenvolvimento de sistemas CAD para a classificação de anomalias em radiografias e a geração automática de laudos médicos preliminares. O objetivo é fornecer uma visão abrangente das abordagens existentes, com ênfase nas técnicas de *machine learning, deep learning* e IA generativa aplicadas a imagens médicas, particularmente na análise da coluna lombo-sacra e dos pododáctilos.

As subseções a seguir exploram duas vertentes principais da pesquisa. A classificação de imagens radiográficas e a geração automática de laudos médicos. Na Subseção 3.1, é apresentada uma análise detalhada dos estudos focados na classificação de radiografias da coluna lombo-sacra, com destaque para os métodos e resultados alcançados. Em seguida, na Subseção 3.2, são discutidos os trabalhos voltados para a geração automática de laudos médicos, com ênfase nas técnicas de IA generativa aplicadas à interpretação de imagens médicas.

Esta revisão foi realizada a partir de pesquisas nas bases de dados *Scopus, Web of Science e IEEE Xplore*, utilizando termos específicos para cada área de estudo, com o objetivo de identificar as abordagens mais relevantes e inovadoras no campo.

3.1 Classificação de Anomalias em Lombo-Sacra

O desenvolvimento de sistemas CAD para a detecção de problemas na coluna vertebral, especificamente na região lombo-sacral, apresenta diversos desafios. A Tabela 3 resume as principais características dos trabalhos relacionados selecionados a partir das bases de dados *Scopus, Web of Science e IEEE Xplore*, sem restrições de período específico, a fim de compreender tanto a evolução da área quanto mitigar possíveis limitações na quantidade de trabalhos disponíveis. Os termos de busca utilizados incluíram *“lumbosacral image classification”, “X-ray image classification” e “lumbar spine with machine learning”*. O critério de seleção foi a aplicação de técnicas de *machine learning e deep learning* para a classificação de anomalias em imagens de raio-X da coluna.

A Tabela 3 resume os principais trabalhos relacionados ao desenvolvimento de sistemas CAD para a coluna vertebral. Antes de destacar os diferenciais de nossa abordagem, é importante analisar os pontos fortes e fracos desses trabalhos existentes.

O estudo de Anna et al. (2019) aborda um problema frequentemente associado a exames da coluna vertebral a escoliose, uma condição. Da mesma forma, Weng et al.

Tabela 3 – Resumo dos trabalhos selecionados na revisão de estado da arte.

Trabalhos	Exame	Incidência	Problema	PP	AD	AP
Cherukuri et al. (2004)	714	lateral	Osteoartrite Osteoporose Osteófitos	sim	não	não
Veronezi et al. (2011)	466	lateral	Osteoartrite primária	não	não	sim
Anna et al. (2019)	595	frontal	Escoliose	não	não	sim
Yaling et al. (2019)	248	frontal	Escoliose	sim	não	sim
Weng et al. (2019)	1,089	lateral	Deslocamento Sagital	não	não	sim
Lee et al. (2019)	334	lateral	Densidade óssea	não	não	sim
Korez, Putzier e Vrtovec (2020)	55	lateral	Balanço Sagitopélvico	não	sim	sim
Schwartz et al. (2021)	816	lateral	Sagitopélvico	sim	sim	sim
Kim et al. (2022)	31,149	frontal lateral	Hérnia de Disco Lombar	não	sim	sim
Naguib et al. (2023)	2,009	lateral	Fraturas Cervical	não	sim	sim
Shen et al. (2023)	12,673	lateral	Fraturas Vertebrais Osteoporóticas	não	não	sim
Liu et al. (2024)	936	lateral	Espondilolistese	sim	não	sim
Zhang et al. (2024)	942	lateral	Fraturas Vertebrais Osteoporóticas	sim	sim	sim

PP: Pré-processamento; AD: Aumento de Dados; AP: Aprendizado Profundo;

(2019) investiga o desvio da coluna, um tópico relacionado. No entanto, Cherukuri et al. (2004) amplia o foco para condições mais complexas que compartilham causas comuns, como osteoartrite, osteoporose e osteófitos, oferecendo uma análise mais abrangente das patologias da coluna. Por sua vez, Veronezi et al. (2011) também explora a osteoartrite, mas especificamente a osteoartrite primária, abordando uma variante distinta da condição.

Os trabalhos de Anna et al. (2019), Yaling et al. (2019), Weng et al. (2019) e Lee et al. (2019), junto com Veronezi et al. (2011), são pioneiros na aplicação de técnicas de aprendizado profundo para a classificação de patologias em imagens da CV. Esses estudos não apenas demonstraram a eficácia do *deep learning* na identificação e categorização de condições patológicas, mas também sugeriram que essas metodologias poderiam ser amplamente aplicadas para melhorar a precisão dos diagnósticos nesta área.

A abordagem descrita em Korez, Putzier e Vrtovec (2020) utiliza aprendizado profundo para medir o balanço sagital espinopélvico, destacando-se pela automação das medições, o que proporciona economia de tempo e redução de erros humanos. Em contraste, Schwartz et al. (2021) concentra-se na automação da medição de parâmetros espinopélvicos em radiografias laterais lombares, utilizando DL para oferecer uma análise mais eficiente e objetiva. Já o estudo de Kim et al. (2022) foca na detecção de hérnia de disco lombar, uma condição de relevância clínica significativa. A aplicação de técnicas de *deep learning* neste estudo resulta em uma melhoria substancial na precisão diagnóstica, evidenciando o potencial das tecnologias avançadas para aprimorar os diagnósticos médicos.

Apesar das contribuições, os trabalhos revisados apresentam algumas limitações. A falta de integração multimodal é uma das principais limitações. Muitos estudos, como Yaling et al. (2019) e Anna et al. (2019), concentram-se exclusivamente em radiografias frontais, enquanto outros, como Cherukuri et al. (2004), Veronezi et al. (2011), Weng et al. (2019), Lee et al. (2019), Korez, Putzier e Vrtovec (2020) e Schwartz et al. (2021), utilizam apenas radiografias laterais. Esses estudos não exploram a combinação de ambas as incidências, o que poderia potencialmente melhorar a análise diagnóstica, fornecendo uma visão mais abrangente das condições patológicas.

Outro ponto crítico é o volume de dados limitado. Trabalhos como os de Korez, Putzier e Vrtovec (2020), Yaling et al. (2019) e Veronezi et al. (2011) utilizam conjuntos de dados relativamente pequenos. Esse fator pode comprometer a generalização e a robustez dos modelos desenvolvidos, tornando-os mais suscetíveis ao *overfitting* e limitando sua capacidade de desempenho em novos conjuntos de dados. Além disso, a escassez de dados restringe a aplicabilidade dos resultados para diferentes populações e cenários clínicos, o que pode reduzir a eficácia dos modelos em situações reais e diversas, onde a variabilidade de dados é maior.

Além disso, a ausência de técnicas avançadas de aumento de dados em muitos dos estudos analisados limita a representatividade dos dados. Técnicas de aumento de dados poderiam melhorar a diversidade dos conjuntos de dados e, consequentemente, o desempenho dos modelos, como observado em Cherukuri et al. (2004) e Liu et al. (2024).

A maior parte dos estudos revisados (dez dos treze) concentra-se em problemas específicos, limitando o escopo das suas abordagens. Por exemplo, trabalhos como o de Anna et al. (2019) concentram-se em escoliose, enquanto Weng et al. (2019) aborda desvio sagital e Lee et al. (2019) foca em densidade óssea. Essa abordagem limitada não cobre a ampla gama de condições da coluna vertebral, o que poderia restringir a aplicabilidade dos modelos desenvolvidos para diagnósticos mais abrangentes e variados.

A falta de abordagens “ponto-a-ponto”, ou seja, metodologias que integrem todas as etapas do pipeline de análise de imagem — desde o pré-processamento até a

classificação — de maneira coordenada, é uma limitação notável. Muitos estudos se concentram em apenas uma etapa, como a classificação, sem considerar a integração das etapas em um fluxo contínuo. A ausência dessa abordagem integrada pode comprometer a precisão e a eficácia dos modelos, já que a coordenação entre todas as fases do processamento de imagens é essencial para garantir diagnósticos mais precisos e completos.

Os estudos também revelam insuficiência em técnicas de *fine-tuning* em modelos pré-treinados. Técnicas avançadas de *fine-tuning* poderiam aprimorar significativamente o desempenho dos modelos, como sugerido pelos trabalhos de Yaling et al. (2019) e Schwartz et al. (2021). A falta dessa prática pode limitar a capacidade dos modelos de se adaptar e melhorar com base em dados específicos.

Finalmente, há uma pouca atenção ao pré-processamento das imagens em muitos dos trabalhos revisados. O pré-processamento é etapa importante para garantir a qualidade dos resultados obtidos e, quando negligenciado, pode impactar negativamente os resultados, como evidenciado nos estudos de Anna et al. (2019) e Weng et al. (2019).

Ao identificar as fraquezas nos trabalhos existentes, fica claro que há oportunidades significativas para melhorias e inovações na área de classificação e detecção de anormalidades em radiografias da coluna vertebral. Nosso trabalho visa abordar essas lacunas e oferecer uma abordagem mais abrangente e robusta, inspirando-se em estudos anteriores, mas oferecendo avanços metodológicos.

Embora os trabalhos relacionados, como os de Liu et al. (2024), Yaling et al. (2019) e Weng et al. (2019), apresentem contribuições valiosas para a classificação de imagens da CV humana, nosso trabalho se diferencia por sua abordagem inovadora e abrangente. A seguir, destacamos os principais diferenciais de nossa pesquisa, com base na literatura revisada.

Nossa abordagem adota uma análise abrangente das condições da CV, considerando uma ampla variedade de patologias, ao contrário de estudos anteriores, como Anna et al. (2019) e Lee et al. (2019), que se concentram em patologias específicas, como escoliose e densidade óssea, respectivamente. Esse enfoque holístico proporciona uma visão detalhada de diversas condições que afetam a coluna, destacando-se como um dos principais diferenciais do nosso trabalho.

Para garantir a robustez do modelo e uma cobertura abrangente de cenários clínicos, compilamos um conjunto de dados composto por 16.024 exames. Esse volume significativo de dados, em contraste com estudos como Korez, Putzier e Vrtovec (2020), que utilizou 55 exames, e Yaling et al. (2019), que utilizou apenas 248 exames, contribui substancialmente para a eficácia e confiabilidade das análises realizadas, proporcionando maior generalização e robustez aos modelos desenvolvidos.

Adotamos um sistema CAD “ponto-a-ponto”, conforme sugerido por Schwartz et al. (2021) em seus estudos de parâmetros espinopélvicos, que integra um fluxo de controle de imagens lombo-sacras. Este sistema é composto por classificação baseada em transferência de aprendizado (YALING et al., 2019), processamento específico e segmentação de marcadores metálicos, conforme recomendado por (ZECH et al., 2018). Essa abordagem criteriosa no pré-processamento das imagens é aplicada tanto no conjunto de treinamento quanto em novas imagens utilizadas durante a inferência, visando garantir resultados mais precisos e confiáveis, refletindo um cuidado detalhado com cada etapa do processamento, como visto também em (WENG et al., 2019).

Além disso, nossa metodologia inclui técnicas avançadas de pré-processamento e aumento de dados. As técnicas de pré-processamento, inspiradas em trabalhos como Ronneberger, P.Fischer e Brox (2015) e Gonzalez (2011), visam aprimorar as características das imagens, enquanto o data augmentation, conforme explorado por Connor (2019), enriquece o conjunto de dados, aumentando sua robustez e diversidade de amostras.

No que diz respeito ao *fine-tuning*, adotamos uma metodologia de *fine-tuning* profundo nas CNNs pré-treinadas (SIMONYAN; ZISSERMAN, 2014), (SIMONYAN; ZISSERMAN, 2015), conforme sugerido por Weng et al. (2019). Essa abordagem permite que o modelo aprenda características profundas das imagens, o que contribui para obter resultados mais robustos e precisos, como demonstrado por Korez, Putzier e Vrtovec (2020) e Schwartz et al. (2021).

Implementamos também uma estratégia de classificação em conjunto (ensemble), semelhante à descrita por Kim et al. (2022), que utiliza imagens de diferentes ângulos do mesmo exame. Combinamos e comparamos os resultados de três arquiteturas distintas de CNN, aproveitando a combinação das informações para melhorar a precisão da classificação.

Por fim, desenvolvemos uma abordagem integrativa baseada em limiares de confiança, como explorado por Vogado et al. (2022). Essa estratégia permite inferir, com maior precisão, a presença ou ausência de anomalias em um exame, integrando informações de múltiplas incidências e modelos para aumentar a robustez e a precisão do diagnóstico.

Os diferenciais de nosso trabalho mostram que estamos abordando as limitações dos estudos existentes e avançando no desenvolvimento de um sistema CAD mais robusto, preciso e abrangente para a detecção de anormalidades na coluna vertebral.

3.2 Geração Automática de Laudos Preliminares

Ao explorar a viabilidade de desenvolver um sistema CAD para a geração automática de laudos médicos preliminares, com foco específico na descrição de achados em radiografias da coluna lombo-sacra e pododáctilos, realizamos uma revisão abrangente da literatura existente. A Tabela 4 resume as principais características dos estudos identificados sobre o uso de IA generativa em imagens médicas. Nossa revisão incluiu pesquisas nas bases de dados *Scopus, Web of Science e IEEE Xplore*, utilizando termos de busca como *"Diagnostic Subtitling", "Automatic Medical Reports", "Medical Reports with Deep Learning", "Medical Reports with Machine Learning", "Automatic Generation of Lumbar Spine Medical Reports with Generative AI" e "Automatic Generation of Pododactyl Medical Reports with Generative AI"*. O critério de seleção principal foi a aplicação de técnicas de aprendizado profundo e IA generativa para a detecção de anomalias em radiografias da coluna vertebral e pododáctilos, bem como outros tipos de imagens médicas relevantes.

A Tabela 4 apresenta um panorama dos estudos que abordam a geração automática de laudos médicos preliminares, evidenciando as diferentes metodologias aplicadas para essa finalidade. Esses trabalhos demonstram o potencial da IA generativa na descrição de achados radiográficos e na produção de relatórios médicos automatizados.

Tabela 4 – Resumo dos trabalhos selecionados na revisão do estado da arte.

Trabalhos	Exames	Imagens	Laudos	Incidências	Problemas	IPP	TPP	AD	Metodologia
Wang et al. (2018)	Raios-X de Tórax	3.643	3.643	Uma	Todos Achados	Não	Sim	Não	Cnns *Long Short-Term Memory*
Huang et al. (2019)	Raios-X de Tórax	Não Informado	Não Informado	Uma	Todos Achados	Não	Não	Não	Multi-Atenção e Incorporação de Informações de Fundo
Kougia et al. (2021)	Raios-X de Tórax	40.306	40.306	Duas	Todos Achados	Não	Sim	Não	CNNs Transformers
Cao et al. (2023)	Endoscopia GI	15.345	3.069	Cinco	Todos Achados	Sim	Sim	Não	Rede de Memória Multi-Modal
Zhao et al. (2023)	Raios-X de Tórax	206,563	206,563	Uma	Todos Achados	Não	Sim	Não	CNN *Transformer*
Mohsan et al. (2023)	Raios-X de Tórax	3.950	3.950	Uma	Todos Achados	Sim	Sim	Não	CNN *Transformer*
Xue et al. (2024)	Câncer de Bexiga	Não Informado	Não Informado	Não Informado	Câncer de Bexiga	Não	Não	Não	Mapeamento Multimodal
Tsaniya, Fatichah e Suciati (2024)	Raios-X de Tórax	9.199	3.973	Uma	Todos Achados	Sim	Não	Não	CNN *Transformer*
Shaik e Cherukuri (2024)	Imagem Retinal	15.709	15.709	Uma	Todos Achados	Sim	Não	Não	CNN *Transformer*
Kong et al. (2024)	Tomografia Computadorizada do Cérebro	10.368	10.368	Fatias	Uma	Sim	Sim	Não	CNN GPT-2

IPP: Pré-processamento de Imagem; PPT: Pré-processamento de Texto;
AD: Aumento de Dados; DL: *Deep Learning*

Para facilitar a compreensão do estado da arte, elaboramos um resumo que não apenas analisa os pontos fortes e fracos dos trabalhos revisados, mas também ressalta como nossa pesquisa se diferencia das abordagens existentes. A análise dos estudos relacionados revelou contribuições notáveis e práticas exemplares no campo da geração automática de laudos médicos preliminares. A seguir, destacamos os principais aspectos que sobressaíram durante a revisão:

O estudo apresentado em Tsaniya, Fatichah e Suciati (2024) evidencia um desempenho aprimorado em métricas de validação de seu modelo, evidenciando a eficácia da abordagem proposta para a geração automática de laudos médicos preliminares a partir de radiografias. De maneira semelhante, Shaik e Cherukuri (2024), apresentou resultados promissores, reforçando a competitividade e a viabilidade das suas técnicas.

Adicionalmente, o estudo de Zhao et al. (2023) destacou-se ao integrar dados médicos complementares, como históricos clínicos e contexto diagnóstico, para aprimorar a qualidade dos laudos gerados. Essa estratégia de enriquecimento de relatórios foi também explorada em Kong et al. (2024).

Estudos como os de Shaik e Cherukuri (2024) e Kong et al. (2024) realizaram experimentos extensivos em diversos conjuntos de dados, demonstrando a viabilidade e eficácia de suas abordagens em comparação com métodos avançados. Essa abordagem experimental abrangente reforça a validade e a confiabilidade dos resultados.

Quanto ao uso de modelos pré-treinados e estudos de ablação, Kong et al. (2024) explorou várias CNNs pré-treinadas para identificar os modelos com melhor desempenho. Em contrapartida, Mohsan et al. (2023) destacou-se pela adoção de modelos avançados baseados em *Transformers*, realizando estudos de ablação para selecionar os componentes mais eficazes. Essas abordagens demonstram uma preocupação rigorosa com a otimização do desempenho e a avaliação criteriosa dos componentes utilizados.

A identificação dos pontos fracos nos estudos existentes destaca áreas com potencial para melhorias e avanços tecnológicos. A seguir, ressaltamos os principais aspectos limitantes observados:

Estudos como os de Wang et al. (2018), Cao et al. (2023) e Mohsan et al. (2023) apresentam uma limitação importante como a falta de informações detalhadas sobre os conjuntos de dados utilizados, o que pode comprometer a generalização e a robustez dos resultados. Por exemplo, Cao et al. (2023) utiliza um conjunto de dados relativamente pequeno, com 3.069 exames, enquanto Wang et al. (2018) e Tsaniya, Fatichah e Suciati (2024) empregam apenas 3.643 e 3.973 exames, respectivamente. A escassez de dados e a ausência de informações detalhadas podem resultar em modelos com menor

capacidade de generalização, impactando a confiabilidade dos resultados.

O estudo de Zhao et al. (2023) enfrenta desafios relacionados à generalização do modelo para diferentes patologias ou contextos clínicos devido a possíveis vieses presentes nos dados históricos dos pacientes. Esses vieses, particularmente na distribuição de anomalias nos dados dos históricos, podem limitar a aplicabilidade do modelo em cenários variados e diversificados.

Além disso, estudos, incluindo Cao et al. (2023), Tsaniya, Fatichah e Suciati (2024) e Shaik e Cherukuri (2024), não menciona explicitamente o uso de técnicas de aumento de dados. A ausência dessas técnicas pode limitar a capacidade dos modelos de lidar com casos mais complexos ou raros, prejudicando a robustez e a adaptabilidade dos modelos.

O trabalho de Kong et al. (2024) apresenta desafios relacionados à interpretabilidade, devido à complexidade das arquiteturas utilizadas e à natureza das tomografias computadorizadas. Essa complexidade pode dificultar a compreensão de como o modelo toma decisões, o que é crucial para a aplicação clínica e para a confiança nos resultados.

Por fim, vários estudos, como Tsaniya, Fatichah e Suciati (2024) e Kong et al. (2024), não fornecem informações detalhadas sobre as técnicas de pré-processamento de dados empregadas. A falta de clareza sobre o pré-processamento pode afetar negativamente a qualidade dos dados de entrada e, consequentemente, o desempenho do modelo.

Nosso trabalho apresenta várias inovações e diferenciais em relação aos estudos existentes, os quais são discutidos a seguir:

Em primeiro lugar, utilizamos um novo conjunto de dados específico, composto por radiografias e laudos médicos das regiões lombo-sacra e pododáctilos. A escolha desses dados permite uma abordagem direcionada e especializada, similar ao que foi proposto por Shaik e Cherukuri (2024) com imagens retinais e por Kong et al. (2024) em tomografias computadorizadas. Assim como nesses estudos, a especificidade dos dados aumenta a relevância e a aplicabilidade do método proposto. A particularidade dos dados radiográficos proporciona uma base sólida para o desenvolvimento de modelos ajustados às particularidades dessas regiões anatômicas, o que é crucial para a precisão diagnóstica, conforme demonstrado em Tsaniya, Fatichah e Suciati (2024) e Zhao et al. (2023).

Além disso, implementamos um sistema CAD ponto-a-ponto que gerencia o fluxo de imagens dessas regiões com uma abordagem detalhada. Esse sistema abrange a classificação baseada em transferência de aprendizado, processamento específico e segmentação de marcadores metálicos, de forma semelhante ao trabalho de Mohsan et al. (2023), que integrou CNNs e *Transformers* para aprimorar a análise de raios-X de tórax.

Essa metodologia reflete um cuidado minucioso no pré-processamento das imagens, algo que é amplamente reconhecido como essencial para garantir resultados mais precisos e confiáveis, conforme observado em Wang et al. (2018) e Cao et al. (2023).

Para garantir a robustez do modelo e a eficácia na classificação, utilizamos técnicas avançadas de pré-processamento e aumento de dados, em conformidade com as práticas recomendadas em Mohsan et al. (2023) e Kong et al. (2024). Essas técnicas aprimoram as características das imagens e aumentam a diversidade dos dados, contribuindo significativamente para melhorar a generalização dos modelos, como discutido por Tsaniya, Fatichah e Suciati (2024) e Shaik e Cherukuri (2024). A aplicação dessas estratégias de aumento de dados reforça a capacidade dos modelos de lidar com cenários clínicos variados, especialmente quando a quantidade de dados disponível é limitada ou apresenta grande variabilidade.

Adotamos ainda o *fine-tuning* profundo em CNNs pré-treinadas, o que nos permitiu ajustar o modelo às características específicas das radiografias de lombo-sacra e pododáctilos. O *fine-tuning* profundo refere-se ao processo de liberar todas as camadas da CNN, inclusive os kernels de convolução, para aprender com o novo conjunto de dados, permitindo uma adaptação completa às novas características (VIEIRA et al., 2021). A eficácia dessa abordagem foi previamente demonstrada por Zhao et al. (2023) e Mohsan et al. (2023), que utilizaram *fine-tuning* para melhorar o desempenho dos seus modelos em tarefas de classificação de exames médicos. Essa técnica permite ao modelo adaptar-se integralmente às novas características das imagens, como evidenciado em Kong et al. (2024), resultando em uma melhor performance nas análises clínicas.

Por fim, propomos uma abordagem inovadora para a geração automática de laudos médicos preliminares, integrando múltiplas imagens do mesmo exame. A combinação de IA generativa, aprendizado profundo e *transformers*, como visto nos trabalhos de Zhao et al. (2023) e Tsaniya, Fatichah e Suciati (2024), possibilita que nosso método auxilie os especialistas na análise das radiografias, oferecendo um suporte adicional fundamentado em informações detalhadas e abrangentes. Esta integração de múltiplos ângulos e imagens é fundamental para enriquecer o processo de auxilio ao diagnóstico, como discutido em Shaik e Cherukuri (2024) e Wang et al. (2018), que também utilizaram múltiplas abordagens para refinar seus modelos generativos.

Nosso trabalho se diferencia ao combinar uma abordagem holística e específica para a detecção de anormalidades na coluna vertebral e nos pés. O uso de um grande volume de dados, técnicas avançadas de pré-processamento e aumento de dados, juntamente com a implementação de metodologias de *fine-tuning* e a integração de múltiplas incidências de imagem, agrega valor ao método proposto. Essas inovações visam superar os desafios e limitações identificados em estudos anteriores, avançando significativamente o campo da geração automática de laudos médicos preliminares.

3.3 Considerações Finais

Neste capítulo, revisamos estudos científicos relevantes sobre a classificação de anomalias em imagens de raios-X da CV e sobre a geração automática de laudos médicos preliminares para as regiões lombo-sacra e pododáctilos. Os métodos e resultados desses estudos foram apresentados brevemente, destacando seus pontos fortes e fracos. Através dessa análise, buscamos fornecer um panorama abrangente das abordagens atuais e identificar áreas com potencial para avanços significativos.

Na seção de classificação de anomalias na CV 3.1, observamos que as abordagens utilizando técnicas de *deep learning* têm mostrado eficácia promissora na identificação de diversas condições na CV. Os estudos revisados se destacaram pelo desempenho aprimorado, pela integração de conhecimento médico e pelo uso de modelos pré-treinados. No entanto, também identificamos limitações, como a ausência de detalhes sobre os conjuntos de dados e a falta de utilização de técnicas como aumento de dados.

Na seção sobre a geração automática de laudos médicos preliminares 3.2, os trabalhos revisados apresentaram abordagens inovadoras que integram conhecimento médico com o uso de arquiteturas avançadas, como *Transformers*. Embora esses estudos tenham alcançado progresso significativo, ainda persistem desafios relacionados à interpretabilidade dos modelos e à escassez de informações detalhadas sobre o pré-processamento dos dados e a restrição dos conjuntos utilizados.

Nosso trabalho se diferencia dos estudos do estado da arte ao empregar conjuntos de dados específicos e extensos, focando em uma ampla variedade de achados nas regiões da coluna vertebral e dos pododáctilos. Implementamos um sistema CAD ponto-a-ponto que utiliza técnicas avançadas de pré-processamento e aumento de dados, além de aplicar *fine-tuning* profundo em modelos pré-treinados. Um diferencial importante é nossa metodologia de classificação em conjunto, que integra imagens frontais e laterais, utilizando um limiar de confiança para inferir com precisão se um exame é normal ou anormal. Além disso, propomos um sistema para geração automática de laudos médicos preliminares que fornece uma valiosa segunda opinião aos especialistas, agregando valor ao processo diagnóstico.

No próximo capítulo, serão apresentados os conceitos teóricos fundamentais para o desenvolvimento deste trabalho. Esses conceitos fornecerão a base necessária para a compreensão das técnicas e metodologias utilizadas, bem como para a implementação e validação do sistema proposto.

4 Classificação de Anomalias em Radiografias da Coluna Lombo-Sacra

Neste capítulo, é apresentado o método para a classificação automática de radiografias da coluna lombo-sacra, com foco na distinção entre imagens normais e anormais. O objetivo central é desenvolver um sistema CAD para auxiliar radiologistas na identificação de anomalias na região lombo-sacra da coluna vertebral, oferecendo suporte adicional no diagnóstico de condições patológicas. Para atingir esse objetivo, propomos uma abordagem que examina as imagens em busca de características relacionadas a diferentes tipos de anomalias. A Figura 13 ilustra o processo. A metodologia completa, desde a preparação dos dados até a avaliação dos resultados, é descrita nas seções subsequentes.

4.1 Método Proposto

A classificação de imagens de radiografias da coluna lombo-sacra foi estruturada em várias etapas visando a precisão e a robustez do modelo final. As etapas principais incluem a aquisição de dados, pré-processamento das imagens, aumento de dados (*data augmentation*), treinamento das CNNs e a classificação utilizando um *ensemble* de modelos com seleção de limiares de confiança.

O processo inicia-se com a aquisição de um conjunto abrangente de radiografias da coluna lombo-sacra. Em seguida, essas imagens passam por um processo de triagem, sendo classificadas com base na coloração dos ossos e no tipo de incidência radiográfica. Na primeira etapa de pré-processamento, as imagens são submetidas a operações de corte, preenchimento com zeros, redimensionamento e segmentação de artefatos metálicos, utilizando a CNN U-Net, preparando-as adequadamente para serem processadas pelas redes convolucionais. Em seguida, o pré-processamento foca na remoção de ruídos e no aprimoramento do contraste das imagens, garantindo uma melhor qualidade visual para o treinamento dos modelos.

Posteriormente, aplicamos técnicas de aumento de dados e ajuste fino em três arquiteturas de CNN (VGG16, VGG19 e ResNet50). Essas redes foram combinadas em um modelo ensemble, utilizando uma metodologia de confiança baseada em limiares. Por fim, os resultados obtidos são analisados e discutidos, avaliando a precisão e confiabilidade do sistema proposto.

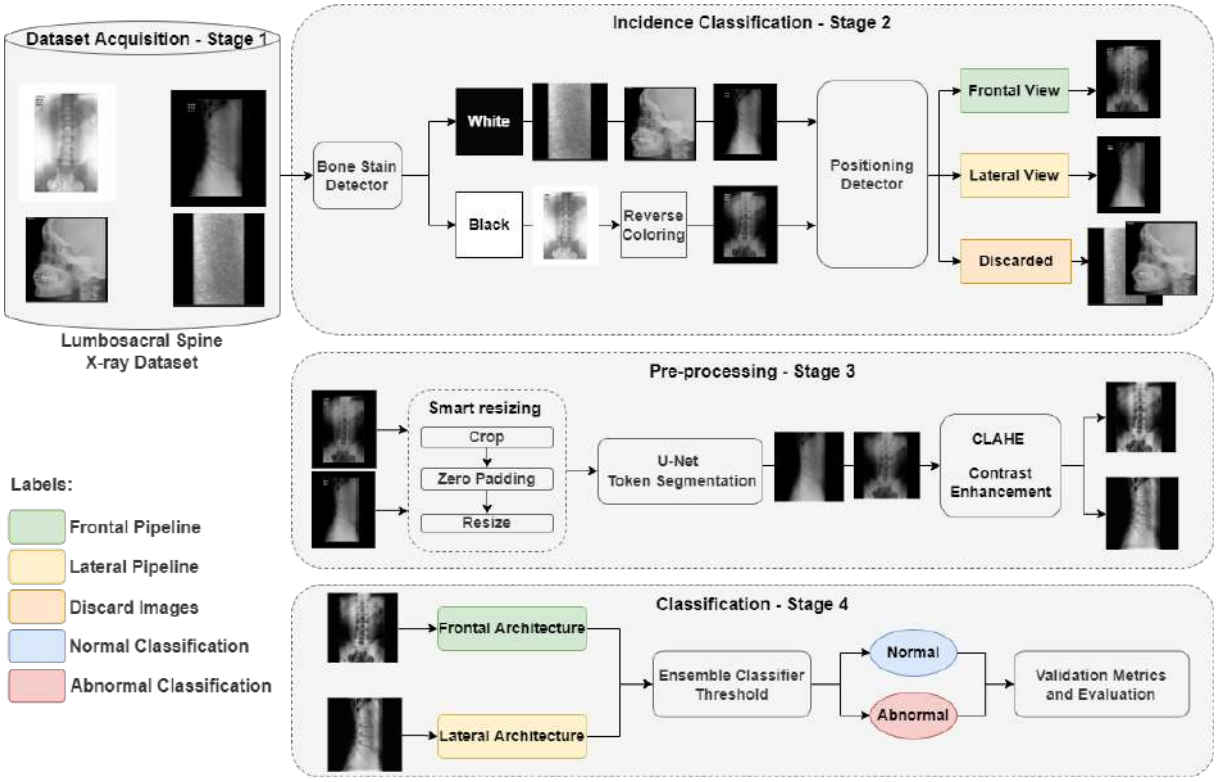


Figura 13 – Fluxograma da metodologia proposta para classificação de radiografias da coluna lombo-sacra.

4.1.1 Aquisição de Imagens

A obtenção de um conjunto de dados robusto e heterogêneo representa um dos maiores desafios no desenvolvimento de sistemas CAD. Para este estudo, reunimos um conjunto de dados composto por 16.024 exames de radiografias da coluna lombo-sacra. Desses exames, 8.265 foram classificados como normais e 7.759 como anormais, sendo as anomalias identificadas a partir de qualquer alteração apontada por especialistas. Cada exame pode incluir duas incidências — uma frontal e uma lateral — resultando em um total de 16.024 imagens frontais e 15.703 laterais. As imagens possuem dimensões originais de até 2140x1760 pixels.

As imagens de radiografias da coluna lombo-sacra apresentam duas incidências principais frontal e lateral, cada uma fornecendo diferentes informações sobre a anatomia da coluna. A Figura 1 ilustra exemplos dessas incidências, destacando as regiões sacral e coccígea, bem como os metadados hospitalares representados por *tokens* metálicos.

Os exames foram coletados em parceria com hospitais e clínicas de todas as regiões do Brasil (Norte, Nordeste, Centro-Oeste, Sudeste e Sul), garantindo uma diversidade significativa em termos de resolução das imagens, tipos de enquadramento, além de diferentes gêneros biológicos e faixas etárias dos pacientes. A classe anormal inclui patologias como desvio ou escoliose da coluna, artrose, espinha bífida,

espondilartrose, lordose, osteófitos, redução do espaço discal, entre outras. Ou seja, para a classe anormal consideramos qualquer alteração ou patologia identificada pelos especialistas. Já a classe normal abrange exames em que não foram observadas anomalias, representando uma anatomia típica e saudável da coluna lombo-sacra, sendo esses exames essenciais para treinar o modelo a reconhecer a normalidade e diferenciar adequadamente dos casos patológicos.

A rotulação dos exames foi realizada por uma equipe de médicos especialistas em radiografias, todos com experiência comprovada. O médico mais experiente da equipe coordenou o processo, garantindo a consistência e a qualidade das rotulações. Essa rotulação foi realizada uma única vez, seguindo uma metodologia padronizada de Interpretação de Laudo por Radiologista (RIR) (ÇALLı et al., 2021; VOGADO et al., 2022), na qual os especialistas analisaram os laudos médicos e classificaram os exames de acordo com seu conteúdo.

O atributo fotométrico[1] das imagens foi configurado com o parâmetro Monochrome2[2], no qual pixels de maior intensidade representam áreas mais claras, e pixels de menor intensidade, áreas mais escuras. A Figura 1 apresenta exemplos de imagens com incidências frontal e lateral, pertencentes ao mesmo exame, destacando as informações complementares fornecidas por cada uma delas. Para garantir a consistência e uniformidade dos atributos das imagens, o formato DICOM[3] foi utilizado. O uso do DICOM tenta padroniza as características das imagens, como resolução, profundidade de bits e fotometria, facilitando a interoperabilidade entre diferentes sistemas e equipamentos médicos.

4.1.2 Triagem das Imagens

A triagem das imagens é uma etapa essencial no desenvolvimento de pipelines para sistemas CAD, assegurando a integridade e relevância dos dados processados. No caso das radiografias da coluna lombo-sacra, é fundamental padronizar aspectos como a coloração das imagens e garantir que as imagens correspondam adequadamente ao tipo de exame e ao ângulo correto. Durante a análise inicial das imagens de lombo-sacra, identificamos inconsistências, onde os metadados do DICOM indicavam a coloração dos ossos de uma forma, mas ao examinarmos as imagens, percebíamos o contrário. Além disso, verificamos a presença de imagens provenientes de outros exames ou sem

[1] O atributo fotométrico refere-se à característica da imagem que determina como os níveis de cinza são interpretados, permitindo diferenciar áreas claras e escuras.

[2] Monochrome2 é um parâmetro que define a escala de intensidade dos pixels em imagens médicas DICOM, onde pixels de maior valor indicam áreas mais claras e pixels de menor valor representam áreas mais escuras.

[3] DICOM (*Digital Imaging and Communications in Medicine*) é o padrão internacional para a gestão, armazenamento, transmissão e visualização de imagens médicas, assegurando interoperabilidade entre sistemas e equipamentos médicos.

informações relevantes da região de lombo-sacra. Essas observações destacaram a necessidade de realizar uma triagem de posicionamento, separando as imagens em incidências frontais, laterais e imagens para descarte. A triagem foi implementada em duas etapas principais para atender a essas demandas.

Classificação da coloração dos ossos, onde as radiografias usadas por especialistas normalmente apresentam os ossos em branco sobre um fundo preto. Contudo, nosso conjunto de dados inclui exames com essa característica invertida. Para solucionar esse problema, foram treinados classificadores como *Support Vector Machines* (SVM), *Random Forests* (RF), *Multi-Layer Perceptrons* (MLP) e XGBoost utilizando carácteristicas extraídas por CNNs pré-treinadas, como VGG16 e ResNet50, utilizando transferência de aprendizado do ImageNet (KRIZHEVSKY; SUTSKEVER; HINTON, 2017). O papel desses classificadores é identificar se a imagem segue o padrão correto (ossos em branco sobre fundo preto). Caso seja detectada uma inversão desse padrão, o classificador ajusta automaticamente a coloração, garantindo a consistência necessária para o processamento subsequente.

Classificação de incidências, devido à heterogeneidade do nosso conjunto de dados, algumas imagens podem não conter informações relevantes, como fundos totalmente pretos ou exames de regiões anatômicas não relacionadas à coluna lombo-sacra. Para garantir que apenas as imagens corretas sejam processadas, implementamos uma classificação que distingue entre três categorias: incidência frontal, incidência lateral (da coluna lombo-sacra) e imagens para descarte. As imagens de descarte são aquelas que não se enquadram nas duas primeiras categorias. A mesma metodologia utilizada para a classificação da coloração dos ossos foi aplicada aqui, empregando classificadores treinados para realizar essa triagem com precisão.

Essa classificação foi inicialmente desenvolvida a partir da experimentação descrita na Seção 4.1.7, onde o modelo classificador foi treinado e validado. Após a separação definitiva das imagens pelo modelo, a triagem foi revisada manualmente e validada pela nossa equipe, assegurando que as imagens categorizadas estavam corretas e apropriadas para os processos subsequentes.

4.1.3 Pré-processamento de Imagens

O pré-processamento é uma etapa importante, pois serve para uniformizar as características das imagens e melhorar a qualidade do *input* para as CNNs. Implementamos um fluxo de pré-processamento apontado a seguir.

Limiarização, onde utilizamos o método de limiarização de Otsu para realçar as bordas das regiões de interesse nas radiografias. Este método é eficaz para segmentar imagens com base na distribuição de intensidades dos pixels, destacando as bordas e

eliminando áreas não relevantes, o que facilita os modelos de aprendizado focarem apenas em características anatômicas importantes (OTSU, 1979).

Preenchimento com bordas zeros e redimensionamento, para padronizar o formato das imagens, aplicamos preenchimento com zeros, transformando-as em um formato quadrado. Isso é essencial, pois as CNNs usadas no nosso estudo adotam esse formato de entrada padronizado. Em seguida, redimensionamos as imagens para 256×256 pixels. Este redimensionamento é feito sem distorção, devido ao formato quadrado obtido após o preenchimento com zeros, garantindo que as características anatômicas das imagens sejam preservadas.

Segmentação de *tokens* metálicos, durante a produção de radiografias é comum a inclusão de *tokens* metálicos contendo informações sobre o hospital ou clínica, assim como metadados. A presença desses *tokens* pode enviesar o aprendizado dos modelos, conforme evidenciado por (ZECH et al., 2018) e (GEIRHOS et al., 2020). Em testes preliminares, utilizamos métodos mais simples, como o limiar de Otsu e o K-means, mas devido à heterogeneidade dos dados, esses métodos não se mostraram eficientes para a segmentação, principalmente devido às variações em tamanho, localização e disposição dos *tokens*. Por isso, optamos pela arquitetura U-Net, uma CNN desenvolvida especificamente para a segmentação de imagens biomédicas (RONNEBERGER; P.FISCHER; BROX, 2015). A U-Net demonstrou uma maior capacidade de generalização, possibilitando a segmentação eficaz dos *tokens* nas diferentes condições apresentadas nas radiografias.

Método de Marcha Rápida (*Fast Marching Method* - FMM), após a remoção dos *tokens* com a U-Net, utilizamos o FMM proposto em (TELEA, 2004) para corrigir a área segmentada. Esse método leva em consideração as informações dos pixels vizinhos, partindo sempre das bordas da região de interesse. Os pixels da região segmentada são substituídos pela soma ponderada e normalizada dos pixels da vizinhança, garantindo uma transição suave e natural na imagem corrigida.

Equalização de histograma, radiografias podem apresentar baixo contraste, o que pode dificultar a detecção de anomalias (VIEIRA et al., 2021). Para mitigar essas limitações, aplicamos a técnica de equalização de histograma adaptativa limitada por contraste (CLAHE). Esta técnica divide a imagem em blocos e equaliza o histograma de cada região. Se houver ruído, as limitações de contraste impedem sua propagação, resultando em um realce significativo das estruturas presentes na radiografia (PIZER et al., 1990).

A Figura 14 mostra os efeitos dessas etapas de pré-processamento nas radiografias originais da coluna lombo-sacra.

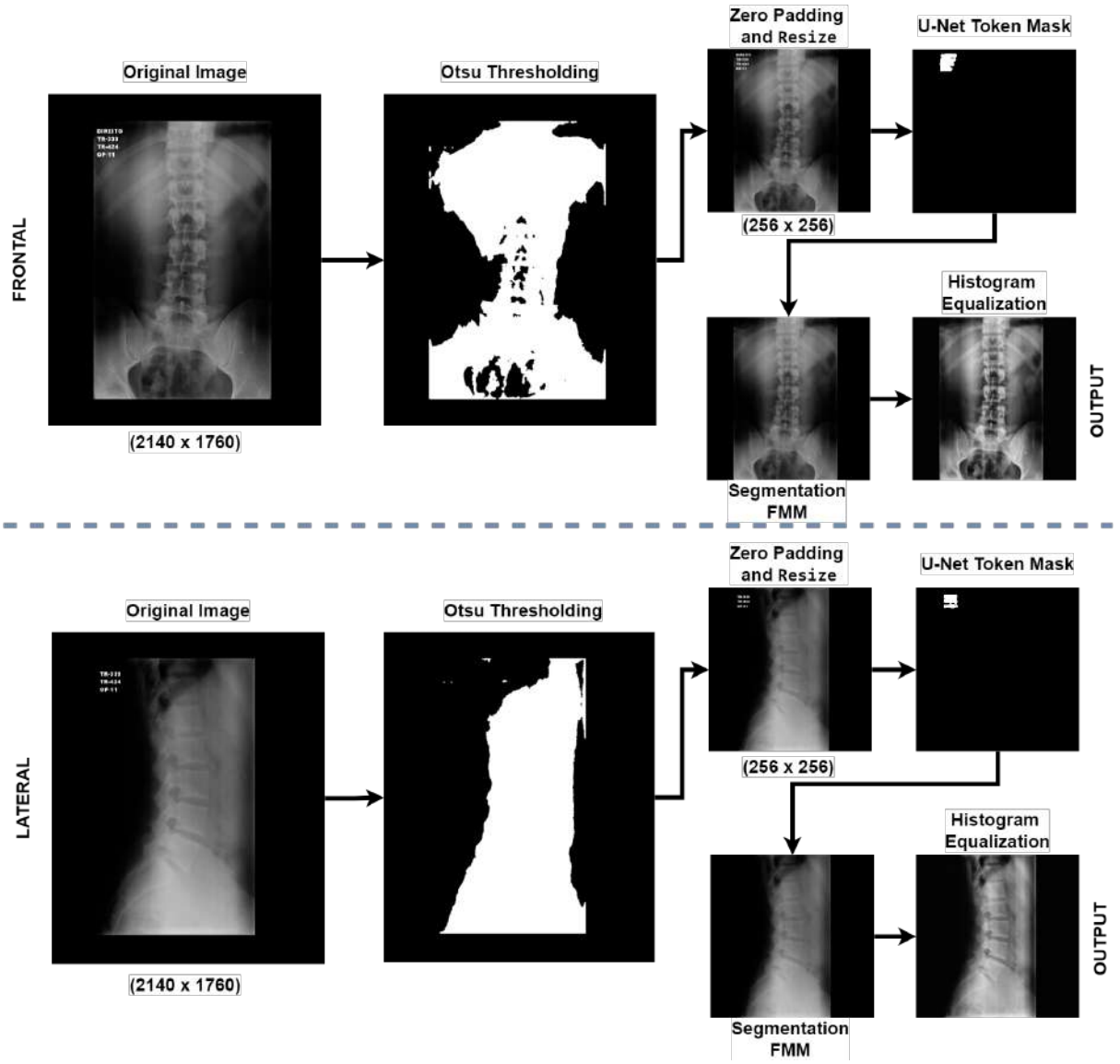


Figura 14 – Exemplo de pré-processamento. frontal e lateral. Para ambas as incidências, adotamos a limiarização de Otsu para remoção de bordas; a adição de preenchimento com zeros para formato quadrado e redimensionamento sem distorção; máscara de segmentação de *tokens* com U-Net; o resultado da segmentação de *tokens* com a correção FMM; e, finalmente, a imagem resultante com a equalização de histograma com o CLAHE.

4.1.4 Aumento de Dados

Para aumentar a variedade das amostras e melhorar o aprendizado das CNNs na tarefa de classificação, utilizamos técnicas de aumento de dados durante o treinamento. As operações incluídas foram rotação de 0º a 10º, alteração de escala em 0% a 5%, deslocamento horizontal e vertical de 2%, espelhamento horizontal, ajuste de brilho de 0% a 20% e adição de ruído gaussiano de 0% a 10%.

Essas operações foram aplicadas de forma aleatória e combinada em tempo de execução, gerando novas imagens sintéticas a cada época de treinamento. Para aplicar o

aumento de dados, utilizamos a biblioteca ImageDataGenerator do Tensorflow, que permite definir a probabilidade de aplicação de cada técnica de aumento em 50%. Com isso, embora a geração variada de amostras seja maximizada, existe a possibilidade, ainda que remota, de não ocorrerem transformações em algumas imagens. Além disso, os parâmetros para cada operação, como rotação, escala e brilho, foram selecionados de forma randômica dentro dos intervalos especificados, proporcionando uma maior diversidade nas amostras ao longo do treinamento. As Figuras 15 e 16 exemplificam essas operações nas imagens frontais e laterais, respectivamente.

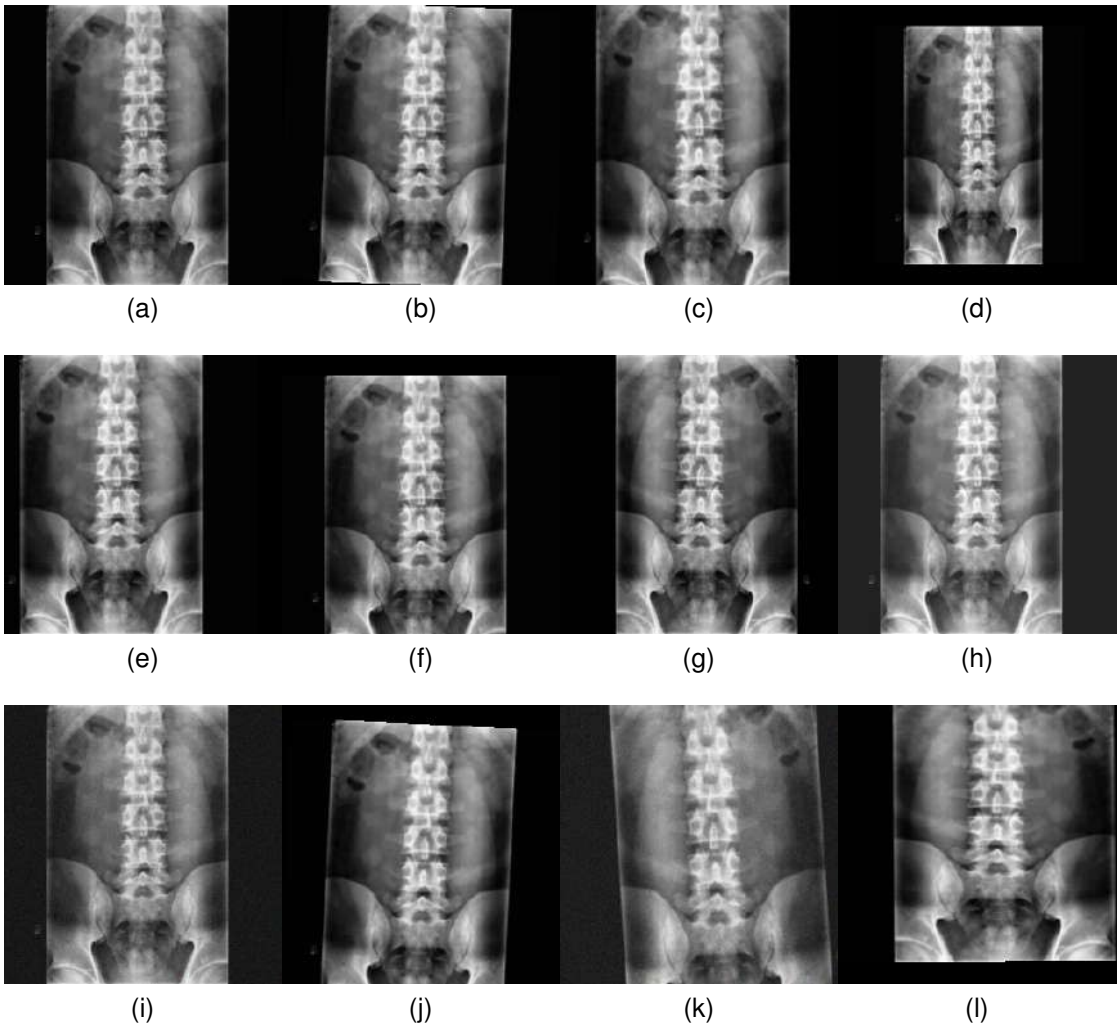


Figura 15 – Exemplos de aumento de dados frontal: (a) Frontal Original; (b) Rotação; (c) Zoom Positivo; (d) Zoom Negativo; (e) Deslocamento Horizontal; (f) Deslocamento Vertical; (g) Espelhamento; (h) Brilho; (i) Ruído Gaussiano; (j, k, l) Combinação Aleatória de Todas as Operações e Suas Faixas.

É importante ressaltar que o conjunto de dados apresenta um desequilíbrio entre as classes normal e anômala, com 51,58% dos dados pertencendo à classe normal e 48,42% à classe anômala. Apesar disso, optamos por não utilizar aumento de dados para balancear o conjunto de dados, mas sim para aumentar a variedade de amostras e maximizar o aprendizado das arquiteturas. Em vez disso, para lidar com o

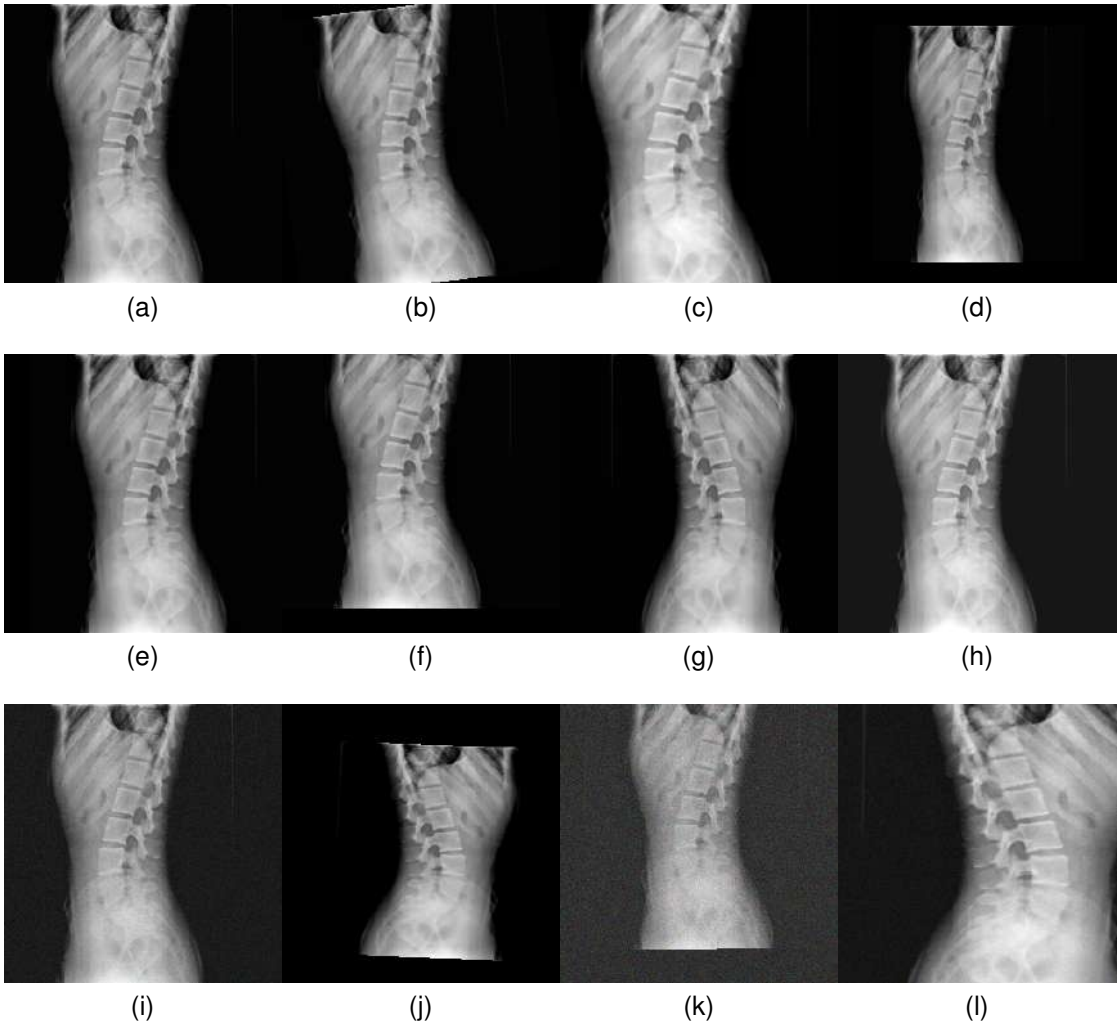


Figura 16 – Exemplos de aumento de dados lateral: (a) Lateral Original; (b) Rotação; (c) Zoom Positivo; (d) Zoom Negativo; (e) Deslocamento Horizontal; (f) Deslocamento Vertical; (g) Espelhamento; (h) Brilho; (i) Ruído Gaussiano; (j, k, l) Combinação Aleatória de Todas as Operações e Suas Faixas.

desbalanceamento, aplicamos a técnica de cálculo de peso das classes (*Compute Class Weight*) (KING; ZENG, 2001).

Na classificação binária, os pesos das classes são calculados com base na frequência das classes positivas e negativas e invertidos para que a classe sub-representada tenha um erro maior do que a classe majoritária quando multiplicado pela perda da classe. A classe menos representada recebe um peso de 1,0 e as outras classes recebem um peso menor que 1,0, baseado na relação com a classe menos representada (KING; ZENG, 2001). Em nosso conjunto de dados, temos 1,0 para a classe anômala e 0,96 para a classe normal.

4.1.5 Treinamento e Classificação

O objetivo desta etapa é realizar o ajuste fino em CNNs previamente treinadas no ImageNet, para classificar as imagens de radiografias da coluna lombo-sacra como anormais ou normais. Utilizamos redes neurais convolucionais (CNNs) previamente treinadas, com foco nas arquiteturas VGG16 (SIMONYAN; ZISSERMAN, 2014), VGG19 (SIMONYAN; ZISSERMAN, 2015), e ResNet50 (HE et al., 2015b). Essas redes foram ajustadas através da técnica de ajuste fino para adaptar-se às características específicas do nosso conjunto de dados. Além disso, foi implementada uma técnica de *ensemble* para combinar as predições das redes, visando aumentar a precisão e confiabilidade dos resultados por meio da definição de um limiar de confiança apropriado.

As arquiteturas VGG16, VGG19 e ResNet50 foram inicialmente pré-treinadas no banco de dados ImageNet(TAJBAKHSH et al., 2016; KRIZHEVSKY; SUTSKEVER; HINTON, 2017). Para adequá-las ao nosso conjunto de dados, realizamos o ajuste fino, substituindo as camadas finais das redes para melhor capturar as particularidades presentes nas radiografias de lombo-sacra. Esse processo permitiu que as redes, que já haviam aprendido a extrair características visuais básicas, se adaptassem a reconhecer padrões mais específicos e relevantes para a detecção de anomalias patológicas no contexto da coluna vertebral.

Nos conjuntos de dados mencionados na Tabela 3, geralmente é utilizada uma única imagem por exame. No entanto, conforme descrito na Seção 4.1.1, nosso conjunto de dados inclui tanto imagens frontais quanto laterais dos exames de lombo-sacra, o que nos permitiu empregar uma abordagem de *ensemble* na classificação entre exames normais e anormais.

A abordagem de *ensemble* foi implementada utilizando duas CNNs distintas, treinadas separadamente com conjuntos de imagens derivados dos diferentes ângulos frontal e lateral. Essa estratégia permite que o sistema capture e analise características anatômicas únicas de cada tipo de visualização, aumentando a robustez da detecção de anomalias ao considerar diferentes perspectivas da coluna lombo-sacra.

Para combinar as predições das diferentes CNNs, adotamos um método de ponderação das saídas individuais das redes no *Ensemble*, integrando-as para gerar um resultado final mais preciso e confiável. O processo de integração é realizado com base em um limiar de confiança pré-estabelecido, que determina a inclusão das predições apenas quando elas apresentam uma alta probabilidade de acerto, descartando aquelas que não atendem aos critérios estabelecidos. Dessa forma, garantimos que a classificação final seja fundamentada em previsões confiáveis, aumentando a precisão e reduzindo a incerteza dos resultados. No entanto, para os casos em que nenhuma das predições atinge o limiar de confiança, os exames são classificados como “dúvida” e

encaminhados ao médico sem uma classificação específica, a fim de evitar a introdução de viés e garantir que os exames sejam revisados cuidadosamente pelo especialista. Essa metodologia oferece uma maior robustez no sistema, garantindo que a classificação final seja respaldada por múltiplas análises independentes, reduzindo o risco de erros e aumentando a acurácia geral do modelo na detecção de anomalias.

4.1.6 Fator de Confiança

Nos sistemas CAD, é importante garantir que as predições fornecidas pelos modelos apresentem confiabilidade, minimizando erros e auxiliando especialistas no diagnóstico de exames complexos. Assim, nosso objetivo é desenvolver um sistema que não apenas classifique amostras corretamente, mas também forneça resultados com um fator de confiança.

Diferentemente dos sistemas tradicionais que utilizam o limiar padrão de 50%, nossa abordagem adota um limiar de confiança mais elevado, idealmente superior a 80%, para assegurar uma maior precisão nas classificações, conforme discutido por Vogado et al. (2022). Ao empregar limiares mais altos, garantimos que as predições, especialmente as da classe normal, sejam feitas com maior certeza, reduzindo a probabilidade de erros.

A Figura 17 ilustra o funcionamento da função de ativação *Softmax* na camada de saída, destacando a seleção dos limiares de confiança. Tradicionalmente, uma amostra é classificada como normal se sua predição ultrapassar o limiar de 50%. No entanto, em nossa metodologia, a classificação só é realizada se a predição atingir os limiares pré-definidos, o que também introduz uma zona de dúvida quando os valores não são alcançados, fornecendo uma indicação de incerteza nas predições.

A função *Softmax* converte o vetor de saídas do modelo em uma distribuição de probabilidade sobre K classes, representando a probabilidade de uma amostra pertencer à classe normal ou anormal (GOODFELLOW; BENGIO; COURVILLE, 2016). Nosso estudo busca definir os melhores limiares de confiança para minimizar falsos positivos, estabelecendo limites mais rigorosos para cada classe. Ao determinar esses limiares por meio da combinação e análise dos resultados, criamos uma zona de dúvida quando o modelo não apresenta alta precisão ou confiabilidade. Nessas situações, optamos por não interferir na análise do especialista, evitando viés na sua percepção. A Figura 18 ilustra o *ensemble* e os limiares definidos.

Para garantir que o *ensemble* funcione corretamente, implementamos regras específicas que asseguram a robustez do sistema. Por exemplo, uma amostra é considerada anormal se, pelo menos, uma das predições (frontal ou lateral) atingir o limiar para anomalia. Por outro lado, para uma amostra ser classificada como normal, ambas as predições devem ultrapassar os limiares definidos. Adicionalmente, quando o exame não

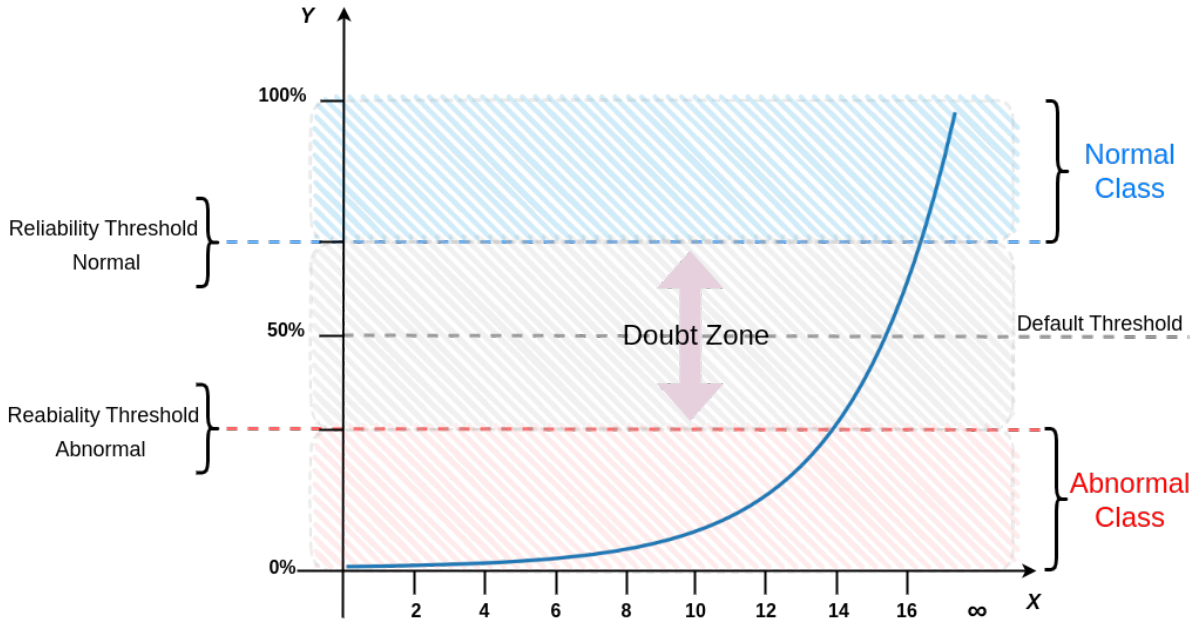


Figura 17 – Representação gráfica da função de ativação *Softmax* com limiares de confiabilidade selecionados.

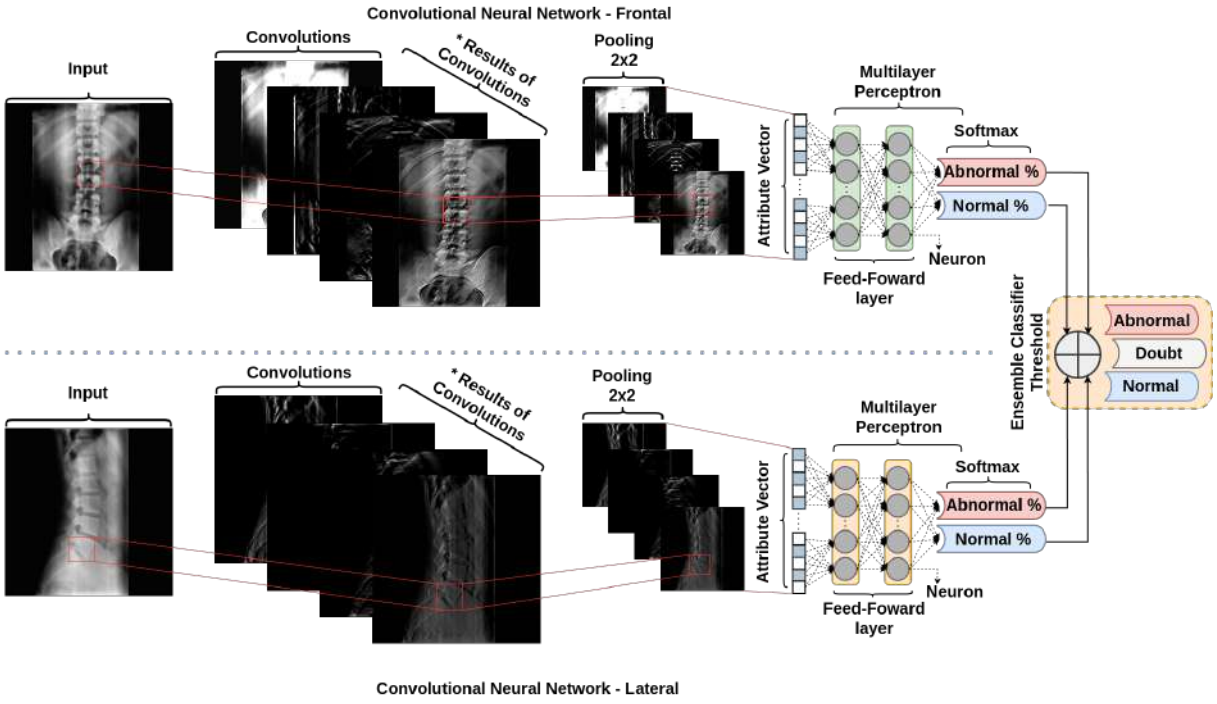


Figura 18 – *Ensemble* proposto composto por duas CNNs especializadas utilizando limiares. No topo, temos uma CNN especializada em imagens frontais. Na parte inferior, temos a CNN especializada em imagens laterais.

contém ambas as incidências (frontal e lateral), a decisão é tomada com base apenas na imagem disponível. Essas regras foram desenvolvidas em conjunto com especialistas da área, conforme a Equação 4.1.

$$C_e = \begin{cases} Abnormal & \text{se ao menos um } CF \leq RT_F^a \text{ ou } CL \leq RT_L^a, \\ Normal & \text{se } CF \geq RT_F^n \text{ e } CL \geq RT_L^n, \\ Doubt & \text{Caso contrário.} \end{cases} \tag{4.1}$$

onde C_e representa o exame a ser classificado; CF representa a predição da imagem frontal e CL da imagem lateral; RT_F^a, RT_L^a, RT_F^n e RT_L^n são os limiares de confiança para frontal anormal, lateral anormal, frontal normal e lateral normal, respectivamente. Além da classificação em anormal e normal, também podemos gerar um estado de "dúvida" quando o resultado dos classificadores não atinge os limiares predefinidos. Nesses casos, o exame é encaminhado para o médico sem uma classificação automática, devendo ser analisado diretamente pelo especialista para garantir uma avaliação precisa e sem introdução de viés.

Para avaliar o desempenho dos modelos, excluímos do cálculo das métricas de validação todos os casos classificados como dúvida, pois, em um ambiente real, esses exames não receberiam classificação automatizada e seriam direcionados diretamente para o especialista. Assim, geramos uma nova matriz de confusão, composta apenas pelos exames classificados com alta confiança, sejam eles normais ou anormais. Essa matriz, sem os casos de dúvida, é então utilizada para calcular métricas como acurácia, $kappa$, entre outras, assegurando que os resultados reflitam exclusivamente as classificações mais seguras.

Quando uma das incidências do exame não está disponível para análise pelos classificadores, nossa proposta é gerar uma $Exceção$ (Equação 4.2).

$$Exception = \begin{cases} \text{if } Frontal\ \emptyset; Frontal \equiv Lateral, \\ \text{if } Lateral\ \emptyset; Lateral \equiv Frontal, \end{cases} \tag{4.2}$$

onde, $\emptyset$ indica que não há amostra de imagem disponível, o que pode ocorrer tanto para o contexto frontal quanto para o lateral. Nesse caso, a incidência ausente é substituída por outra incidência com o resultado equivalente ao classificador dessa incidência. Ou seja, o resultado de uma única imagem é replicado.

4.1.7 Experimentos

Neste trabalho, os dados foram divididos em conjuntos de treinamento, validação e teste, com percentuais de 75%, 10% e 15%, respectivamente. A escolha desses valores se deve ao objetivo de maximizar a quantidade de dados disponíveis para o treinamento do modelo, enquanto ainda reservamos uma proporção adequada para validação e teste.

O conjunto de treinamento (75%) é usado para ajustar os parâmetros do modelo, o conjunto de validação (10%) é utilizado para ajustar hiperparâmetros e evitar sobreajuste durante o desenvolvimento, e o conjunto de teste (15%) para medir o desempenho do modelo em dados que ele nunca viu, garantindo uma avaliação realista de sua capacidade de generalização. Essa distribuição, detalhada na Tabela 5, equilibra o uso eficiente dos recursos computacionais e a robustez na avaliação do modelo.

Tabela 5 – Descrição das tarefas, classes, amostras e métodos utilizados no desenho experimental deste trabalho.

Tarefa	**Classes**	**Amostras por Classe**	**Método**
Coloração do Osso	preto branco	4,421 4,421	TL
Detecção de Incidência	frontal lateral descarte	1,278 1,262 1,881	TL
Segmentação de Tokens	normal anormal	500 500	U-Net
Classificação Frontal	normal anormal	8,265 7,759	FT
Classificação Lateral	normal anormal	4,183 4,454	FT

TL: *Transfer-learning*; FT: *Fine-tuning*.

Classificação da coloração dos ossos, especialistas geralmente utilizam radiografias com a coloração dos ossos em fundo branco e preto. Em nosso conjunto de dados, alguns exames podem apresentar essa característica invertida. Para resolver esse problema, treinamos classificadores de ponta (SVM, RF, MLP e XgBoost) com características extraídas pelas CNNs VGG16 e ResNet50 pré-treinadas no conjunto ImageNet através de aprendizado por transferência, onde geramos um classificador auxiliar que determina se a imagem está ou não na cor padrão dos ossos em branco e o fundo preto, caso contrário, a inversão é realizada.

Detecção de incidência, nosso conjunto de dados é heterogêneo e contém algumas imagens sem informações relevantes, por exemplo, apenas um fundo preto. Em alguns casos, exames diferentes da região lombo-sacra são enviados por engano. Uma parte importante da nossa metodologia é encontrar as imagens corretas para o sistema; nesta etapa, procuramos três tipos de imagens frontal, lateral e descarte [4]. A imagem de descarte pode ser qualquer imagem que não se enquadre nos critérios das duas primeiras classes. Para isso, utilizamos a mesma metodologia do item anterior.

Segmentação de *tokens*, conforme mencionado anteriormente, o conjunto de dados utilizado neste trabalho apresenta marcações de *tokens* metálicos nas radiografias. Isso nos levou a propor a segmentação dos *tokens* usando uma U-Net, que apresenta

[4] Tudo aquilo que não se enquadra em imagens frontais ou laterais de raiografias de lombo-sacra.

bom desempenho na segmentação de imagens médicas. Para esse propósito, foram selecionadas 500 imagens frontais e 500 laterais, e os especialistas produziram as marcações. Essa quantidade foi escolhida para garantir uma amostra representativa dos diferentes padrões de *tokens* metálicos encontrados no conjunto de dados, mantendo o equilíbrio entre a diversidade das amostras e a viabilidade de rotulação manual pelos especialistas, considerando os recursos de tempo e esforço disponíveis. Além disso, para garantir uma melhor generalização da rede, dividimos as imagens em classes normais e anormais, com metade das imagens para cada classe.

Classificação em normal ou anormal, nossa metodologia recebe as radiografias do exame lombo-sacral. Após algumas etapas, deve informar se encontrou uma anomalia. Para a detecção de anomalias, utilizamos as CNNs ResNet50, VGG16 e VGG19 pré-treinadas no ImageNet, onde aplicamos o ajuste fino profundo. Nosso conjunto de dados nos permite usar imagens frontais e laterais dos pacientes para gerar a classificação e treinar diferentes modelos para imagens frontais e laterais, onde aplicamos um *ensemble* com seleção de limiar para melhor desempenho.

Para este trabalho, utilizamos um ambiente de computação com as seguintes configurações, 187 GB de RAM, CPU com 64 núcleos, e duas GPUs Tesla T4 com 16 GB cada, sistema operacional Linux Red Hat. A programação foi realizada em Python 3, utilizando bibliotecas como TensorFlow, Keras, OpenCV, Sklearn, Numpy, e Skimage.

4.2 Resultados

Nesta seção, discutimos os resultados dos experimentos de classificação para a triagem de imagens, desde a coloração dos ossos até a classificação das incidências. Também apresentamos os resultados das CNNs para a classificação de anomalias em radiografias lombos-sacras frontais e laterais. Por fim, avaliamos o método de *ensemble* com seleção de limiar de confiança para os dois contextos de posicionamento.

A Tabela 6 mostra os resultados da classificação da coloração dos ossos. A metodologia obteve uma acurácia (Acc) superior a 98% em todas as combinações. A combinação da VGG16 com o classificador MLP alcançou o melhor resultado. Analisando os resultados, podemos ver que, independentemente da CNN e do classificador, os resultados são semelhantes, variando de 0,984 a 0,995 em termos de Acc. Esses resultados apoiam o uso deste procedimento como parte de nossa metodologia, pois alcançam um resultado “quase perfeito” de acordo com a métrica $Kappa$.

A Tabela 7 apresenta os resultados dos experimentos de detecção de incidência. Observamos que a ResNet50 superou a VGG16, alcançando uma acurácia de 100% quando combinada com o classificador SVM. A maior profundidade da ResNet50 permite extrair características mais discriminativas e detalhadas das imagens, capturando

Tabela 6 – Resultados da metodologia proposta obtidos na detecção de cores ósseas.

CNN	Classificador	Acc	$kappa$	Prec	F1-Score	AUC
VGG16	RF	0,984	0,977	0,984	0,988	0,988
	SVM	0,994	0,988	0,993	0,994	0,994
	MLP	**0,995**	**0,990**	**0,997**	**0,995**	**0,995**
	XgBoost	0,990	0,981	0,991	0,991	0,990
ResNet50	RF	0,987	0,975	0,984	0,987	0,987
	SVM	0,994	0,988	0,995	0,994	0,994
	MLP	0,994	0,988	0,995	0,994	0,994
	XgBoost	0,992	0,984	0,993	0,992	0,992

Em **negrito** os melhores resultados.

nuances importantes para a distinção entre incidências. Quando essas características profundas são combinadas com a capacidade dos *kernels* SVM de lidar eficientemente com grandes conjuntos de dados e separar classes mesmo em espaços não lineares, o modelo atinge um desempenho excepcional.

Tabela 7 – Resultados da metodologia proposta obtidos na detecção de posicionamento.

CNN	Classificador	Acc	$kappa$	Prec	F1-Score	AUC
VGG16	RF	0,923	0,896	0,910	0,918	0,941
	SVM	0,973	0,963	0,970	0,972	0,980
	MLP	0,982	0,976	0,979	0,982	0,987
	XgBoost	0,966	0,954	0,960	0,964	0,974
ResNet50	RF	0,973	0,963	0,965	0,969	0,978
	SVM	**1,0**	**1,0**	**1,0**	**1,0**	**1,0**
	MLP	0,993	0,991	0,992	0,993	0,995
	XgBoost	0,986	0,981	0,984	0,985	0,990

Em **negrito** os melhores resultados.

Além disso, é importante destacar que a tarefa de classificar as incidências radiográficas (frontal, lateral ou descarte) não é intrinsecamente complexa, pois há diferenças visuais claras entre essas categorias. As imagens frontais e laterais da coluna lombo-sacra apresentam características anatômicas distintas que podem ser facilmente reconhecidas por modelos bem treinados, enquanto as imagens de descarte diferem significativamente das demais. Portanto, a combinação de uma CNN profunda como a ResNet50 com um classificador SVM é particularmente eficaz para esta tarefa específica.

Os resultados da classificação da coloração dos ossos e dos experimentos de detecção de posicionamento demonstram a eficácia da metodologia proposta de transferência de aprendizado. Resultados acima de 0,990 de acurácia foram alcançados pela extração de características usando CNNs e utilizando essas características como entradas para classificadores. Os experimentos de classificação da coloração dos ossos revelaram que a combinação da rede VGG16 com o classificador MLP foi eficaz, enquanto a rede ResNet50 superou a VGG16 nos experimentos de detecção

de posicionamento. Essas descobertas apoiam o uso da metodologia proposta de transferência de aprendizado como um componente integral da metodologia geral.

Os resultados da detecção de *tokens* metálicos em radiografias de lombo-sacras indicam que o modelo U-Net alcançou uma acurácia de 0,991, sugerindo que a tarefa pode ser realizada com alta precisão. O coeficiente de *Dice*, que mede a sobreposição entre as máscaras previstas e as máscaras de referência, foi de 0,902, o que significa que 90,2% dos pixels na máscara prevista coincidem com as máscaras produzidas por especialistas. Além disso, o índice de Jaccard, que mede a similaridade entre as duas máscaras, foi de 0,806, indicando que 80,6% dos pixels na máscara prevista estão presentes na máscara de referência fornecida pelos especialistas.

As diferenças nos resultados entre as três métricas refletem que cada uma enfatiza aspectos diferentes. Por exemplo, o coeficiente de *Dice* enfatiza a sobreposição entre as máscaras. Em comparação, o índice de Jaccard enfatiza a proporção de pixels na máscara prevista que também estão presentes na verdade-terreno. Por outro lado, a métrica de acurácia para segmentação enfatiza a proporção de pixels corretamente classificados em relação ao número total de pixels na imagem. Portanto, a acurácia considera tanto os pixels corretamente quanto os incorretamente segmentados.

Esses resultados demonstram que a arquitetura U-Net possui capacidade de generalização, pois conseguiu realizar a tarefa com precisão de sobreposição entre as máscaras previstas e as de verdade-terreno.

Nos experimentos de classificação de radiografias de lombo-sacra contendo classes anormais e normais, os resultados dos classificadores treinados em imagens frontais e laterais foram apresentados na Tabela 8. Na primeira seção (Experimentos sem Segmentação e sem CLAHE), os resultados mostraram que a maior acurácia foi obtida pela VGG16 e VGG19 com 0,787 e 0,806, respectivamente, em imagens frontais. Para imagens laterais, a maior acurácia foi obtida pela VGG16 e ResNet50 com 0,816 cada. O maior valor de $kappa$ foi obtido pela VGG16 em imagens laterais com 0,572. A maior especificidade foi obtida pela VGG16 em imagens laterais com 0,836, enquanto a maior sensibilidade foi obtida pela VGG16 em imagens frontais com 0,881. Com base nesses resultados, a VGG16 é a melhor CNN para essa tarefa de classificação, pois alcançou a maior acurácia, $kappa$, especificidade e sensibilidade em ambas as imagens frontais e laterais.

Apesar dos resultados encorajadores apresentados na primeira seção da Tabela 8, o $kappa$ ainda mostra resultados “fracos” abaixo de 0,490 para imagens frontais e inferior a 0,580 para laterais. Dessa forma, uma nova série de experimentos foi realizada, apresentada na seção 2 (Experimentos sem Segmentação e com CLAHE) da Tabela 8. Na segunda série de experimentos sem segmentação e com CLAHE, o contraste foi melhorado, pois as imagens de raios-X geralmente têm baixa qualidade nesse aspecto.

No entanto, os resultados mostraram que, apesar da melhoria na acurácia, tanto os modelos frontais quanto os laterais apresentaram sensibilidade na detecção de anomalias inferior a 0,860, e o $kappa$ apresentou resultados inferiores quando comparado à primeira seção da tabela. Considerando esses resultados individualmente, concluímos que o uso de CLAHE não é necessário.

Tabela 8 – Resultados obtidos na classificação anormal e normal das duas posições.

CNN	Acc	$kappa$	Prec	F1-Score	AUC	*Specificity*	*Sensitivity*
	Experimentos Sem Segmentação e Sem CLAHE						
	Imagens Frontais						
ResNet50	0,724	0,461	0,760	0,796	0,736	0,643	0,819
VGG16	0,787	0,487	0,835	0,786	0,803	0,739	0,881
VGG19	0,806	0,478	0,773	0,790	0,815	0,751	0,812
	Imagens Laterais						
ResNet50	0,816	0,512	0,836	0,808	0,834	0,734	0,774
VGG16	0,816	0,572	0,847	0,857	0,817	0,836	0,802
VGG19	0,805	0,545	0,812	0,846	0,805	0,819	0,773
	Experimentos Sem Segmentação e Com CLAHE						
	Imagens Frontais						
ResNet50	0,794	0,424	0,730	0,744	0,781	0,713	0,787
VGG16	0,810	0,473	0,811	0,764	0,814	0,765	0,856
VGG19	0,819	0,464	0,751	0,768	0,809	0,775	0,789
	Imagens Laterais						
ResNet50	0,835	0,501	0,817	0,790	0,837	0,754	0,757
VGG16	0,845	0,553	0,821	0,832	0,846	0,866	0,777
VGG19	0,834	0,527	0,785	0,820	0,834	0,848	0,748
	Experimentos Com Segmentação e Sem CLAHE						
	Imagens Frontais						
ResNet50	0,788	0,427	0,735	0,750	0,787	0,707	0,793
VGG16	0,805	0,476	0,817	0,769	0,819	0,759	0,862
VGG19	0,812	0,474	0,767	0,784	0,822	0,757	0,805
	Imagens Laterais						
ResNet50	0,836	0,501	0,816	0,789	0,836	0,755	0,756
VGG16	0,837	0,558	0,828	0,839	0,838	0,858	0,784
VGG19	0,827	0,531	0,791	0,826	0,827	0,842	0,753
	Experimentos Com Segmentação e Com CLAHE						
	Imagens Frontais						
ResNet50	0,799	0,597	0,782	0,808	0,798	0,760	0,843
VGG16	0,825	0,650	**0,841**	0,824	0,825	**0,842**	**0,887**
VGG19	**0,829**	**0,658**	0,820	**0,835**	**0,829**	0,808	0,861
	Imagens Laterais						
ResNet50	0,850	0,700	0,844	0,854	0,850	0,834	0,782
VGG16	**0,865**	**0,729**	**0,887**	**0,863**	**0,865**	**0,890**	**0,840**
VGG19	0,851	0,703	0,861	0,852	0,852	0,859	0,820

Em **negrito** os melhores resultados.

Na terceira seção da Tabela 8 (Experimentos com Segmentação e sem CLAHE),

testamos a classificação frontal e lateral de imagens com segmentação de *tokens*. Nessa seção, para ambos os contextos, observamos que obtivemos um $kappa$ superior ao encontrado na seção experimentos sem segmentação e com CLAHE, porém ainda inferior ao da seção experimentos sem segmentação e sem CLAHE. Também é possível observar que a sensibilidade para remoção de *tokens* é maior em comparação ao uso de CLAHE, mas ainda inferior à primeira seção. A acurácia da seção 2 é melhor do que a obtida na seção 3, o que indicou que poderíamos tentar combinar as duas metodologias para encontrar um resultado superior ao uso das imagens sem essas alterações.

Na última seção da Tabela 8, combinamos a técnica de segmentação com o CLAHE para melhorar a qualidade das imagens no treinamento das CNNs, obtendo resultados superiores em relação a outras abordagens apresentadas na mesma tabela. A acurácia superou 80% para quase todas as arquiteturas, com $kappa$ acima de 0,65 para imagens frontais e 0,72 para laterais, indicando generalização. As CNNs VGG16 e VGG19 apresentaram os melhores resultados, devido à sua menor complexidade e arranjo eficiente de camadas de convolução e *pooling*, o que facilita a generalização e reduz o risco de *overfitting*, especialmente em um cenário com menos amostras e classes, como o contexto lombo-sacro.

A Tabela 9 apresenta o conjunto com a combinação de arquiteturas frontais e laterais; aqui, mostramos apenas os melhores resultados na seleção de limiares para classificação. Inicialmente, consideramos o intervalo de limiares entre 0,7 e 1,0 para a classe normal e 0 a 0,3 para a classe anormal. Assim, fizemos combinações dos limiares (variando todas as combinações possíveis de seus intervalos) para encontrar uma taxa de classificação com menos falsos positivos. A razão para escolher os valores de variação indicados acima é que as combinações desses limiares com as redes de cada incidência têm um alto custo computacional, então limitamos o espaço de busca. Aqui, não mostramos todos os resultados obtidos para cada combinação de limiar possível, pois a quantidade de dados impressos dificultaria sua interpretação.

A VGG-16 apresentou o melhor resultado entre a taxa de intervenção dos modelos e o erro de normais, com uma intervenção de 22,31% e um FDR de 2,79% (cometendo um erro no pior caso, FP igual a dez exames), conforme apresentado na Tabela 9. A VGG-16 teve o menor erro entre as combinações avaliadas. Considerando a taxa de intervenção com o erro para outras arquiteturas, como ResNet50, VGG-16/VGG-19, VGG-16/ResNet50 e ResNet50/VGG-16, a VGG-16 alcançou resultados superiores. Na Tabela 9, destacamos a porcentagem de respostas da VGG-16 em comparação com as combinações com erros semelhantes. Considerando a arquitetura VGG-19 para classificação frontal, observamos uma alta taxa de intervenção relacionada a um alto erro. Portanto, os valores de limiar dessa arquitetura são altos, o que não permite a redução de erros.

Tabela 9 – Tabela de resultados obtidos na classificação em anormal e normal, nas imagens frontal e lateral, utilizando o conjunto com seleção de limiares das arquiteturas de classificação de anomalias para incidências.

Thresholds								
Frontal	Lateral	MC	Dúvidas	qtd Respostas	Normal	FDR	Anormal	FOR
	Experimentos com Modelos VGG16 (Frontal) & VGG16 (Lateral).							
0,1 0,7	0,1 0,94	**349 10** **49 586**	615 38,22%	**994** **61,78%**	**22,31%** (359)	**2,79%** (10)	39% (635)	7,72% (49)
	Experimentos com Modelos VGG19 (Frontal) & VGG19 (Lateral).							
0,03 0,99	0,03 0,82	543 46 73 640	307 19,08%	1302 80,92%	37% (589)	7,81% (46)	44% (713)	10,24% (73)
	Experimentos com Modelos ResNet50 (Frontal) & ResNet50 (Lateral).							
0,15 0,72	0,15 0,93	335 10 43 576	645 40,09%	964 59,91%	21% (345)	2,9% (10)	38% (619)	6,95% (43)
	Experimentos com Modelos VGG16 (Frontal) & VGG19 (Lateral).							
0,1 0,72	0,09 0,97	314 11 50 595	639 39,71%	970 60,29%	20% (325)	3,38% (11)	40% (645)	7,75% (50)
	Experimentos com Modelos VGG19 (Frontal) & VGG16 (Lateral).							
0,01 0,99	0,01 0,92	490 40 57 587	435 27,04%	1174 72,96%	33% (530)	7,55% (40)	40% (644)	8,85% (57)
	Experimentos com Modelos VGG16 (Frontal) & ResNet50 (Lateral).							
0,09 0,85	0,09 0,96	314 9 46 568	672 41,77%	937 58,23%	20% (323)	2,79% (9)	38% (614)	7,49% (46)
	Experimentos com Modelos ResNet50 (Frontal) & VGG16 (Lateral).							
0,15 0,7	0,15 0,92	314 9 40 589	657 40,83%	952 59,17%	20% (323)	2,79% (9)	39% (629)	6,36% (40)
	Experimentos com Modelos VGG19 (Frontal) & ResNet50 (Lateral).							
0,02 0,99	0,02 0,9	540 45 67 613	344 21,38%	1265 78,62%	36% (585)	7,69% (45)	42% (680)	9,85% (67)
	Experimentos com Modelos ResNet50 (Frontal) & VGG19 (Lateral).							
0,15 0,7	0,15 0,93	331 10 42 592	634 39,4%	975 60,6%	21% (341)	2,93% (10)	39% (634)	6,62% (42)

Em **negrito** os melhores resultados; Matriz de Confusão (MC),
FDR: *False Discovery Rate*; FOR: *False Omission Rate*.

A seleção de limiares de confiança para o *ensemble* visa reduzir a taxa de falsos positivos; para isso, gera uma margem de incerteza. No entanto, como mostrado na Tabela 10, a metodologia melhorou significativamente a classificação, conforme ilustrado pelas métricas de avaliação. Ao analisar as métricas, observamos uma variação positiva de 14,06% e 8,79% na acurácia, 34,46% e 19,89% no $kappa$, 15,44% e 9,21% na especificidade, e 10,82% e 17,02% na sensibilidade quando comparados aos contextos frontal e lateral, respectivamente. Essa variação positiva demonstra que a metodologia *ensemble* com seleção de limiar de confiança aprimora a confiabilidade das classificações ao combinar informações complementares das incidências frontal e lateral, resultando em maior robustez e precisão do modelo em comparação ao uso isolado das CNNs.

A utilização da metodologia *ensemble* com seleção de limiar de confiança possibilita a detecção mais precisa de patologias que podem ser visíveis apenas em uma das incidências, frontal ou lateral, mas não necessariamente em ambas. Por exemplo, condições como osteófitos (bico de papagaio), que são frequentemente mais visíveis

Tabela 10 – Comparação dos melhores resultados para lateral e frontal e com a aplicação da metodologia de ensemble com seleção de limiar de confiança.

CNN	Acc	$kappa$	Prec	F1-Score	AUC	*Specificity*	*Sensitivity*
	Imagens Frontais						
VGG16	0,825	0,650	0,841	0,824	0,825	0,842	0,887
	Imagens Laterais						
VGG16	0,865	0,729	0,887	0,863	0,865	0,890	0,840
	Imagens Frontais & Laterais						
Ensemble	**0,941**	**0,874**	**0,983**	**0,952**	**0,947**	**0,972**	**0,983**

Em **negrito** os melhores resultados.

na incidência lateral devido à sua projeção óssea, podem passar despercebidas na incidência frontal. Ao combinar as informações de ambas as incidências, o *ensemble* garante uma cobertura diagnóstica mais completa, aumentando a robustez do modelo na detecção de anomalias. Dessa forma, o uso dessa abordagem contribui para reduzir a chance de falsos negativos e melhora a precisão geral do diagnóstico, cobrindo melhor as variabilidades inerentes à visualização de diferentes patologias em diferentes ângulos.

4.2.1 Discussão

As CNNs são conhecidas como caixas-pretas, apesar de apresentarem resultados robustos. Nesse contexto, Selvaraju et al. (2016) desenvolveu um método para apoiar a explicação dos resultados. Esse método, chamado *Gradient-weighted Class Activation Mapping* (Grad-CAM), permite visualizar as regiões importantes para a classificação das CNNs. As Figuras19 e 20 apresentam essas regiões para imagens frontais e laterais, indicando que as arquiteturas conseguiram se especializar nas incidências apresentadas durante o treinamento.

A Figura 19a mostra como a CNN interpretou uma imagem frontal, encontrando o que o especialista relatou como uma redução na densidade óssea, osteófitos em vários corpos vertebrais, redução de vários espaços intervertebrais e espondilólise com espondilolistese de L4 sobre L5. A Figura 19b ilustra que a CNN encontrou uma redução da densidade óssea e osteófitos em vários corpos vertebrais e uma redução de vários espaços intervertebrais, conforme relatado pelo especialista. A Figura 19c mostra os pontos de Grad-CAM frontal que coincidem com o relatório do especialista indicando desvio do eixo lombar para a direita, osteófitos em vários corpos vertebrais e redução de vários espaços intervertebrais. A Figura 19d mostra que os achados da CNN coincidem com o relatório do especialista com exame de controle pós-operatório de artrodese na coluna lombar com fixação com placa metálica e parafusos de L4 a S1, osteófitos anteriores incipientes, redução do espaço discal de L4 a S1 e esclerose interapofisária de L5-S1.

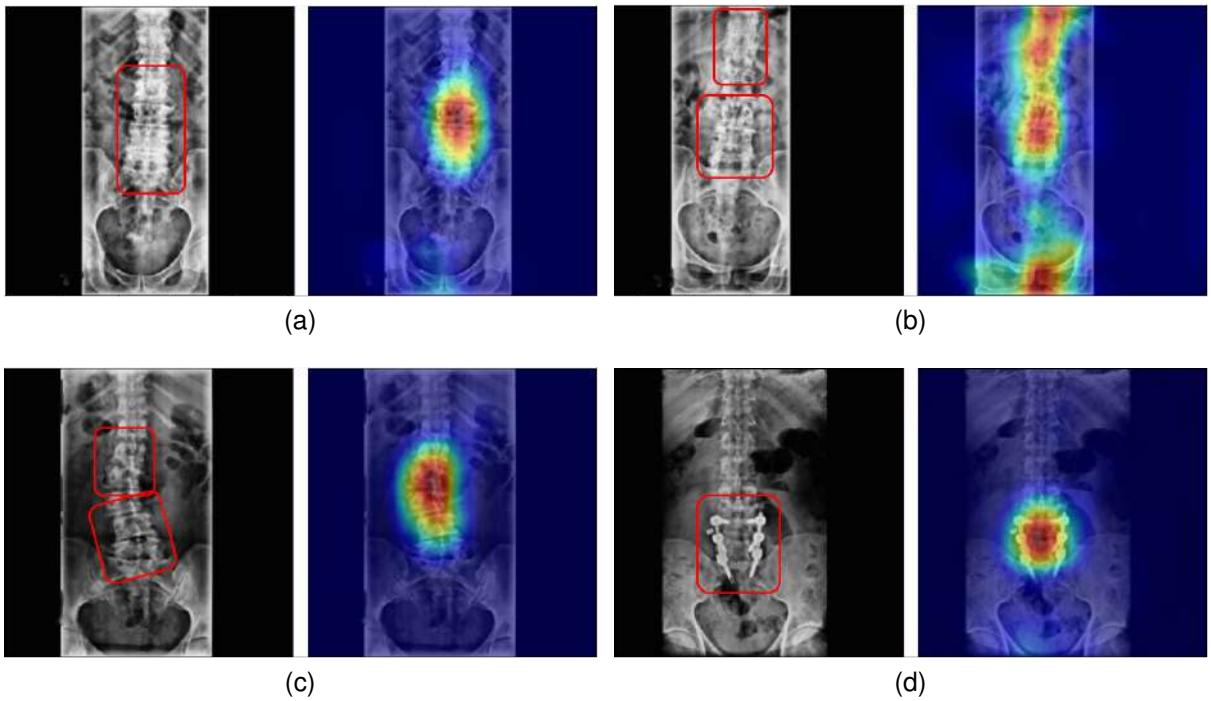


Figura 19 – Exemplos de Grad-CAM para imagens frontais. De (a) a (d), observamos a radiografia do paciente processada por nossa metodologia e, à direita, o resultado visual do Grad-CAM para cada radiografia. A marcação dos principais achados específicos do especialista está destacada em um quadro vermelho.

A Figura 20a mostra como a CNN interpretou uma imagem lateral, correspondendo ao relatório do especialista indicando alterações nos contornos dos corpos vertebrais de L3 a S1 pela presença de formações de osteófitos marginais, redução do espaço intervertebral L5-S1. A Figura 20b mostra que a CNN corresponde ao relatório do especialista com escoliose lombar com convexidade à esquerda, osteófitos em vários corpos vertebrais e redução de vários espaços intervertebrais. A Figura 20c apresenta o Grad-CAM de uma imagem lateral, que aponta para o relatório do especialista aler osteófitos anteriores e laterais de L1 a L5 com tendência a formar sindesmófitos, osteopenia com redução de altura das vértebras lombares, redução dos espaços discais L2-L3, L3-L4, L4-L5 e L5-S1 e esclerose interapofisária de L3-L4, L4-L5 e L5-S1. A Figura 20d mostra a precisão da CNN na detecção das lesões relatadas pelo especialista com corpos vertebrais de forma, densidade, estrutura e contornos anatômicos, pedículos, lâminas e apófises transversas sem alterações, redução do espaço discal L5-S1 e presença de fios de Kirschner.

Para entender melhor os resultados dos Grad-CAMs, geramos a Figura 21 para analisar como as melhores CNNs dos dois contextos tratam as imagens para obter suas previsões, desde as transformações ocorridas nos primeiros blocos de convoluções até os últimos. Ela mostra radiografias frontais e laterais sendo processadas, com seus

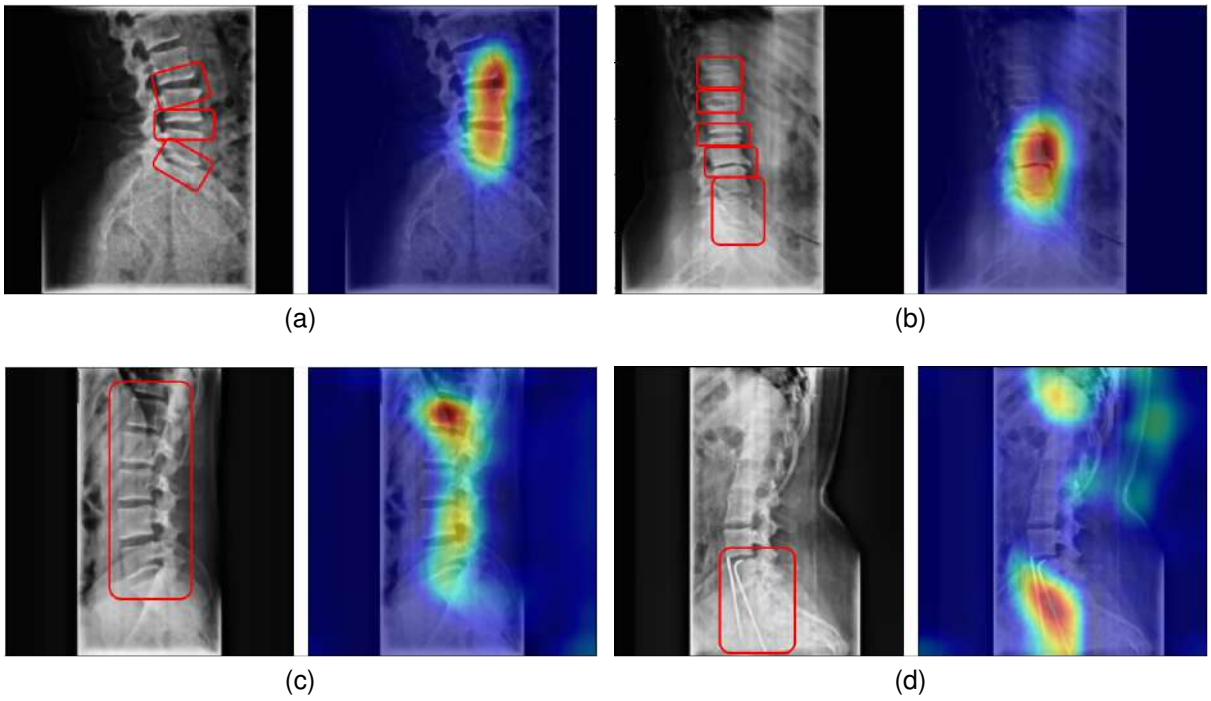


Figura 20 – Exemplos de Grad-CAM para imagens laterais. De (a) a (d) observamos a radiografia do paciente processada por nossa metodologia e, à direita, o resultado visual do Grad-CAM para cada radiografia. A marcação dos principais achados específicos do especialista está destacada em um quadro vermelho.

respectivos Grad-CAMs, pelos blocos de convolução das melhores CNNs apontadas na Tabela 8. A ideia por trás dos blocos de convolução é a possibilidade de usar múltiplas convoluções antes da amostragem. Essa abordagem resolve o problema da rápida diminuição da resolução espacial quando uma camada de convolução é sempre seguida por uma amostragem conforme Andrew e Scott (2016). A Figura 21a ilustra todo o processo de transformação de uma radiografia frontal realizada pela CNN, enquanto a Figura 21b apresenta o processo para uma radiografia lateral.

As Figuras 21a e 21b mostram que os Grad-CAMs gerados nos primeiros e segundos blocos explicam que as CNNs, nesta fase, ainda não especificam uma região de atenção; ainda não temos ativações fortes nesta etapa. O resultado da saída do terceiro bloco mostra que as CNNs começam a apontar para regiões que podem trazer informações importantes para o resultado, focando principalmente nos ossos nas regiões da coluna, demonstrando que estão procurando algo nessa área. Os Grad-CAMs do quarto bloco já começam a prestar mais atenção às regiões que os especialistas apontam em seus relatórios, mas ainda temos alguma dispersão da atenção por parte das ativações. Finalmente, no último bloco, o Grad-CAM indica que as CNNs já conseguiram focar sua atenção exclusivamente na região das anomalias, uma vez que as ativações estão concentradas apenas nessas regiões. Isso só é possível empregando um ajuste

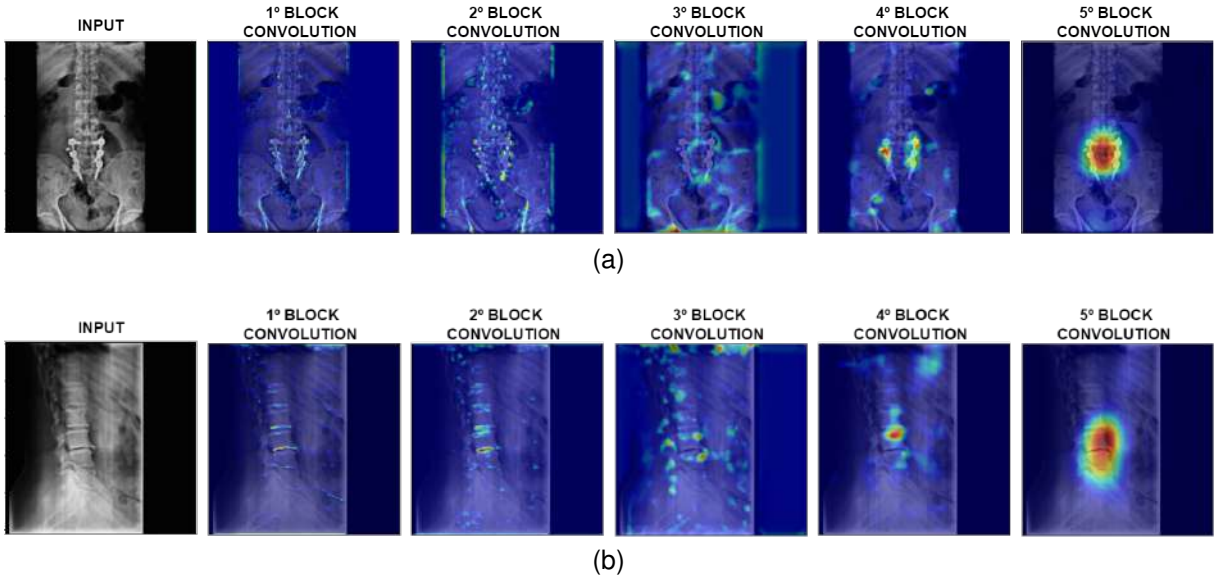


Figura 21 – Exemplos de Grad-CAMs para todos os blocos de convolução das melhores arquiteturas de CNN das ambas posições.

fino profundo, que permite às CNNs aprender características específicas do problema também nas camadas de convolução. Assim, as CNNs aprenderam a detectar lesões em vez de decorar a base de imagens.

Nas classificações, especificamente nas frontais, as CNNs tiveram dificuldade em alcançar bons resultados nos primeiros experimentos. Para entender a razão, investigamos os motivos com especialistas em radiografia. A análise mostrou que algumas doenças eram mais perceptíveis em imagens laterais, como doenças osteofíticas, lordose e redução discal nas vértebras L5-S1. Assim, removemos algumas imagens de treinamento da classificação frontal, com o apoio de especialistas. Dessa forma, o classificador de imagens laterais seria responsável por relatar essas questões para o *ensemble*. No final, o conjunto de dados teve as características apontadas na Seção 4.1.1.

A medicina não é uma ciência exata; alguns exames podem levar a conclusões diferentes pelos especialistas, como percebemos na classificação de desvio de coluna. Nossos dados são de hospitais de todas as regiões do Brasil, produzidos por vários especialistas em condições adversas e dispositivos diferentes, por isso encontramos diferenças na classificação dos exames. Algumas imagens que as CNNs classificaram como contendo desvio foram classificadas por médicos como normais (o oposto também ocorreu). Isso ocorreu ao treinar a CNN com uma imagem que apresentava características de uma classe; no entanto, outra imagem com características semelhantes pertencia a outra classe, impactando o aprendizado do modelo. Para resolver esse problema, reformulamos os relatórios desses exames com a ajuda de médicos especializados em radiografia.

4.3 Conclusões

Neste trabalho, desenvolvemos uma metodologia CAD ponta-a-ponta que integra modelos de inteligência artificial preditiva para a triagem de radiografias da coluna lombo-sacra, com o objetivo de otimizar a identificação de exames que necessitam de atenção prioritária. Nosso objetivo principal foi alcançado ao criar um sistema capaz de auxiliar médicos no processo diagnóstico, melhorando a eficiência e precisão na detecção de anomalias nessa região.

Utilizando um conjunto de dados robusto e diversificado, composto por 16.024 exames, superamos a limitação comum em estudos anteriores que frequentemente utilizam conjuntos de dados menores. As etapas de pré-processamento, incluindo a remoção de ruído e a segmentação de *tokens* metálicos, foram fundamentais para melhorar a uniformidade e qualidade das imagens, eliminando potenciais vieses no aprendizado das CNNs e garantindo maior precisão ao modelo.

Aplicamos técnicas de aumento de dados para enriquecer o conjunto de imagens, permitindo um treinamento mais eficaz e melhorando a capacidade de generalização do modelo. O uso de arquiteturas auxiliares para detectar a coloração óssea e as incidências contribuiu significativamente para a triagem automatizada, aumentando a precisão dos resultados e atendendo ao objetivo de otimizar a identificação e classificação de patologias.

A implementação de CNNs especializadas para imagens frontais e laterais, combinadas com a técnica de *ensemble* e seleção de limiares de confiança, mostrou-se eficaz em aumentar a robustez das classificações e minimizar falsos negativos. Isso permitiu que cada tipo de imagem fornecesse informações adicionais sobre o paciente, contribuindo para um processo diagnóstico mais ágil e eficiente, e com isso possibilitando reduzir a carga de trabalho dos profissionais de saúde, conforme proposto nos objetivos.

No que diz respeito à avaliação do desempenho do modelo, utilizamos métricas específicas e comparações com métodos existentes, validando a precisão, confiabilidade e eficiência do sistema proposto. As métricas alcançadas pelo *ensemble* com limiar de confiança destacam a competitividade e relevância do modelo desenvolvido no contexto atual. Isso atende ao terceiro objetivo específico, assegurando que o modelo não só é eficaz, mas também confiável para aplicação clínica.

Uma limitação significativa do estudo é a utilização de um conjunto de dados privado, o que impossibilita a disponibilização pública dos dados para a comunidade científica. Essa restrição dificulta a replicação dos resultados por outros pesquisadores e limita a comparação direta com trabalhos correlatos. Portanto, futuros estudos poderiam explorar o uso de bases de dados públicas ou estabelecer parcerias que permitam o compartilhamento controlado dos dados, promovendo maior colaboração e avanço na

área.

Como trabalhos futuros, recomendamos a inclusão de novos dados no estudo, visando aprimorar o aprendizado dos modelos e aumentar sua capacidade de generalização. A adição de um número maior de exames, especialmente de casos anormais, tem o potencial de melhorar ainda mais o desempenho do modelo. Além disso, será valioso explorar arquiteturas de redes neurais mais eficientes em termos computacionais, como MobileNet e EfficientNet, bem como redes mais robustas e profundas, como NasNetLarge, para investigar como diferentes arquiteturas podem impactar a precisão e o desempenho do modelo.

Em suma, este estudo alcançou os objetivos propostos, demonstrando que a integração de modelos de IA preditiva em sistemas CAD pode aprimorar a precisão e eficiência diagnóstica na análise de radiografias da coluna lombo-sacra. O trabalho contribui para a redução da carga de trabalho dos profissionais de saúde e fornece suporte valioso na tomada de decisão clínica, promovendo um ambiente clínico mais seguro e eficaz. Esperamos que este trabalho inspire pesquisas futuras e promova avanços adicionais na aplicação da inteligência artificial na prática médica, com o objetivo final de melhorar os resultados para os pacientes.

5 Geração Automática de Laudos Médicos Preliminares em Radiografias de Lombo-Sacra e Pododáctilos

Nesta seção, apresentamos uma aplicação específica do sistema CAD a geração automática de laudos médicos preliminares para radiografias das regiões da coluna e dos pés. A proposta consiste em estender as capacidades do sistema para analisar imagens de lombo-sacara e pododáctilos, identificando características relevantes e gerando descrições textuais precisas e concisas, semelhantes aos laudos produzidos por radiologistas especializados. A metodologia detalhada, incluindo as adaptações necessárias para a análise de radiografias dos pés, será apresentada nas próximas subseções.

5.1 Método Proposto

A metodologia deste trabalho seguiu uma sequência estruturada de etapas. Primeiramente, realizou-se a aquisição do conjunto de dados por meio da coleta de um conjunto abrangente tanto de radiografias de lombo-sacara quanto de pododáctilos. Em seguida, foi feita a triagem das imagens no contexto de lombo-sacra, classificando-as quanto à coloração dos ossos e ao tipo de incidência. O processo continuou com o pré-processamento das imagens, que incluiu corte, preenchimento com zeros, redimensionamento, segmentação de *tokens* metálicos e melhoria do contraste. Paralelamente, foi realizado o pré-processamento dos textos, envolvendo análise, limpeza, processamento, tokenização e geração de *embeddings* dos textos dos laudos. Posteriormente, procedeu-se à extração e interpretação de características, onde as características das imagens foram extraídas utilizando CNNs pré-treinadas, seguidas pela interpretação dessas características com *transformers*. A etapa subsequente envolveu a geração do laudo médico, utilizando modelo generativo para criar automaticamente os laudos médicos. Por fim, foi realizada a análise e discussão dos resultados, avaliando a precisão e confiabilidade do sistema proposto. As Figuras 22, 23 e 26 ilustram as etapas do método para radiografias de lombo-sacra e as Figuras 25 e 26 o processo para radiografias de pododáctilos.

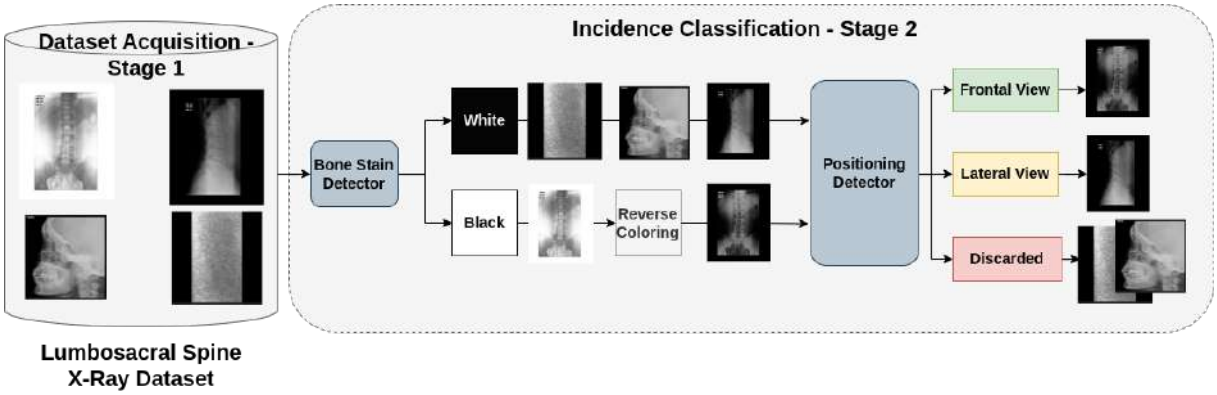


Figura 22 – Fluxograma das etapas aquisição do conjunto de dados e classificação de incidências quanto a coloração dos ossos e tipo de imagem da metodologia proposta para geração de laudos da coluna lombo-sacra.

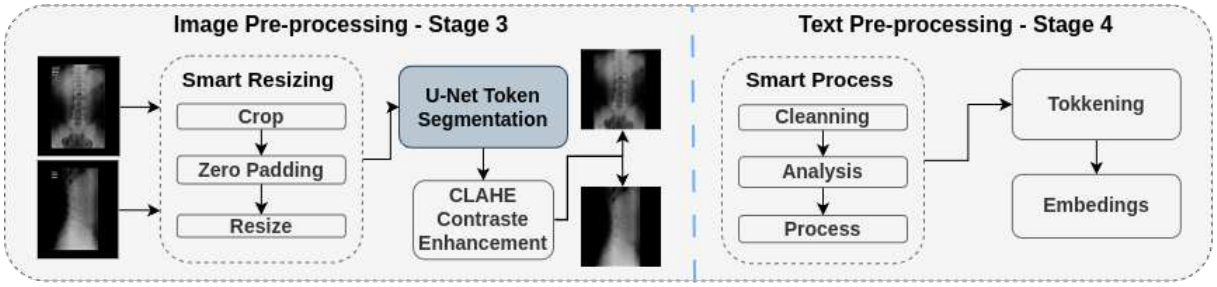


Figura 23 – Fluxograma das etapas de processamento das imagens e dos laudos da metodologia proposta para geração de laudos da coluna lombo-sacra.

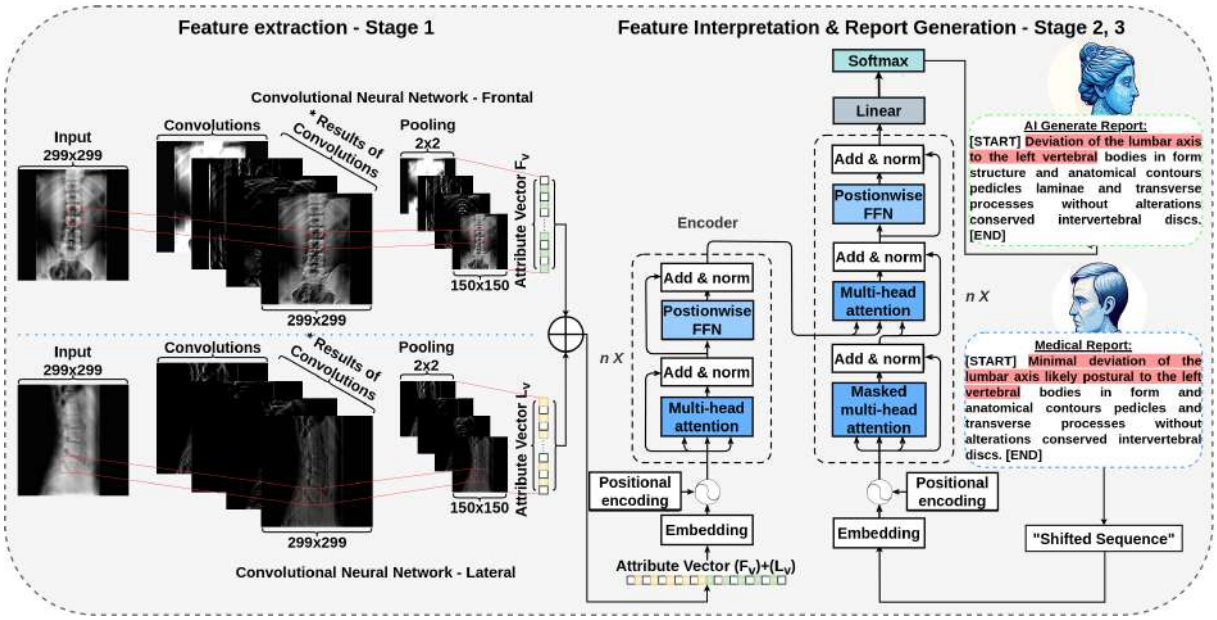


Figura 24 – Ultima etapa da metodologia proposta, geração automática de laudos médicos preliminares a partir de radiografias da coluna lombo-sacra utilizando modelos generativos.

5.1.1 Aquisição de Imagens

Um dos principais desafios no desenvolvimento de sistemas CAD que empregam modelos generativos é a obtenção de conjuntos de dados amplos e heterogêneos,

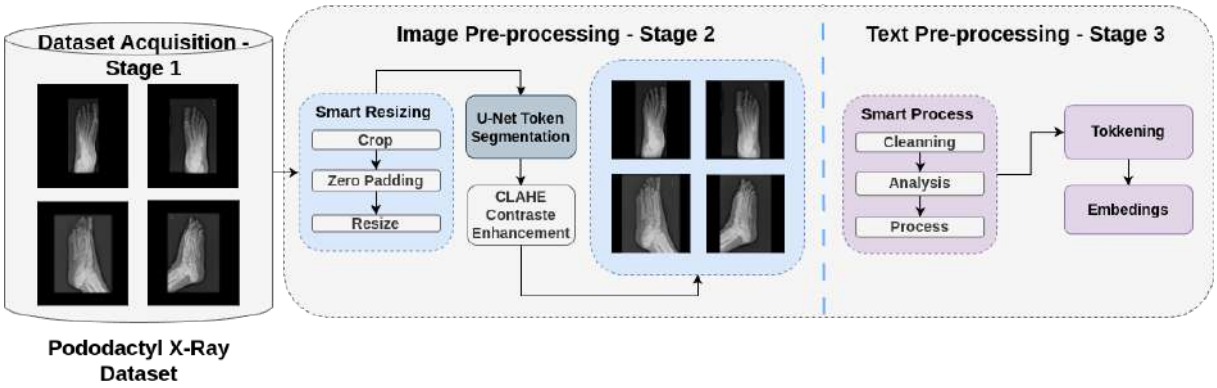


Figura 25 – Fluxograma das etapas aquisição do conjunto de dados e de processamento das imagens e dos laudos da metodologia proposta para geração de laudos de pododáctilos.

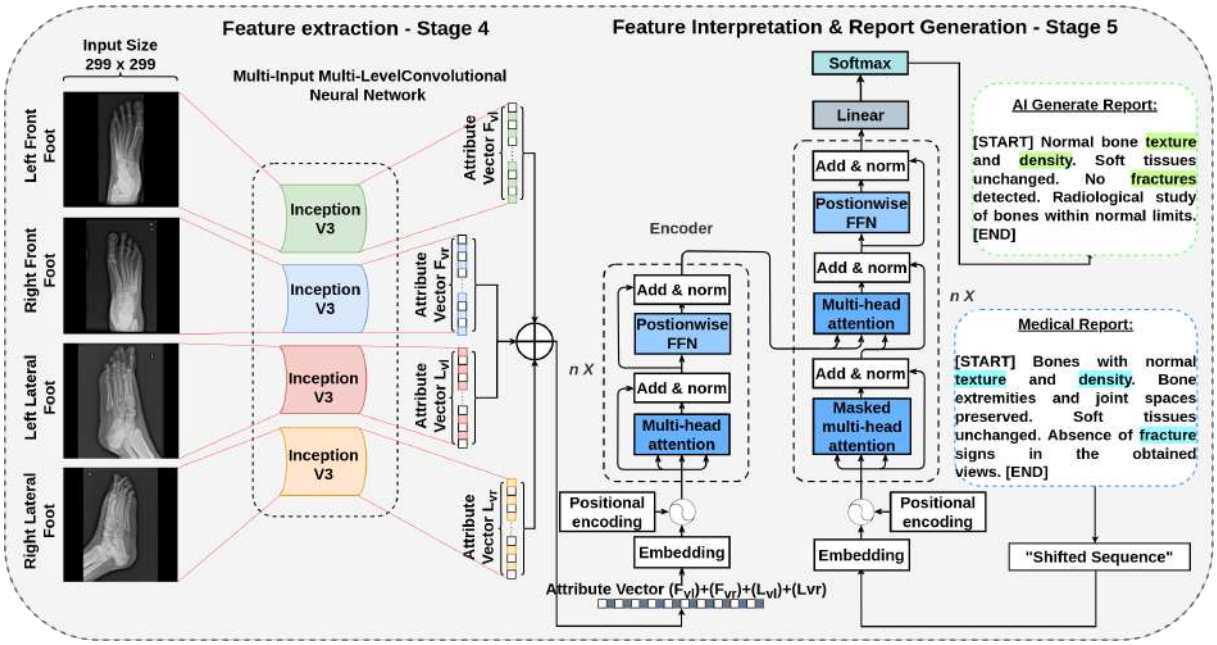


Figura 26 – Ultima etapa da metodologia proposta, geração automática de laudos médicos preliminares a partir de radiografias de pododáctilos utilizando modelos generativos.

fundamentais para o treinamento eficaz dos modelos. Para este estudo, reunimos um conjunto de dados que inclui 44,540 radiografias de lombo-sacra. Destas, 19.702 (44,2%) amostras são classificadas como normais e 24.838 (55,8%) como anormais, como mostrado na Tabela 11. O desequilíbrio observado nesse conjunto, com uma maior quantidade de exames anormais, reflete o uso frequente dessas radiografias para identificar achados clínicos específicos. Embora existam *labels* que separam as classes normais e anormais, o foco principal é o uso dessas radiografias e seus respectivos laudos para gerar relatórios médicos automaticamente, sem a necessidade de classificar os exames.

O conjunto de dados de pododáctilos, também apresentado na Tabela 11, é

Tabela 11 – Dados dos conjuntos lombo-sacra e pododáctilos

Conjunto de Dados	Classe Anormal	Classe Normal	Total	% Anormal	% Normal
Lombo-Sacra	24.838	19.702	44.540	55,8%%	44,2%
Pododáctilos	10.043	6.667	16.710	60,1%	39,9%

composto por 16,710 radiografias, das quais 6,667 (39,9%) são classificadas como normais e 10,043 (60,1%) como anormais. Esse conjunto também apresenta um desequilíbrio, com uma predominância de exames anormais, reforçando a ideia de que as radiografias são mais comumente utilizadas para diagnosticar achados clínicos específicos. Assim como no caso das radiografias lombo-sacras, a classificação entre normal e anormal é considerada apenas para a análise dos resultados, já que o objetivo principal deste estudo é a criação automatizada de laudos médicos a partir dos dados disponíveis.

Embora o desbalanceamento seja maior tanto no conjunto de dados de pododáctilos quanto no de lombo-sacra, com uma maior quantidade de amostras anormais, não consideramos isso um problema significativo no contexto de geração automatizada de laudos médicos preliminares. Entendemos que a maior representatividade das amostras anormais é vantajosa, pois oferece uma maior variabilidade nas características que as definem, permitindo que o modelo capture uma ampla gama de condições anômalas. Por outro lado, os casos normais, sendo mais homogêneos e em menor proporção, não afetam negativamente a capacidade do modelo de identificar padrões anômalos. Portanto, esse desbalanceamento pode contribuir para a robustez e eficácia do sistema na geração de relatórios médicos automatizados, alinhando-se ao nosso objetivo de desenvolver um modelo que cubra uma ampla gama de condições clínicas.

A rotulação dos exames foi realizada por uma equipe de médicos especialistas em radiografias, todos com experiência comprovada. O médico mais experiente da equipe coordenou o processo, garantindo a consistência e a qualidade das rotulações. Essa rotulação foi realizada uma única vez, seguindo uma metodologia padronizada de Interpretação de Laudo por Radiologista (RIR) (ÇALLı et al., 2021; VOGADO et al., 2022), na qual os especialistas analisaram os laudos médicos e classificaram os exames de acordo com seu conteúdo. Para os exames anormais, qualquer característica que fugisse à normalidade foi considerada para a classificação como anomalia, garantindo assim uma avaliação criteriosa e confiável.

Estes conjuntos de dados foram coletados em hospitais e clínicas em todas as regiões do Brasil (Norte, Nordeste, Centro-Oeste, Sudeste e Sul), proporcionando maior heterogeneidade e uma ampla variedade de resoluções, ângulos, sexos e faixas etárias diferentes, totalizando 44,540 e 16,710 amostras de exames com imagens

e laudos anônimos para lombo-sacra e pododáctilos respectivamente.Os exames foram obtidos no formato DICOM e convertidos para o formato ".png" para posterior processamento. O formato ".png" foi escolhido por ser amplamente adotado pela literatura, permitindo compressão sem perdas visíveis, mantendo a integridade necessária para o treinamento e validação dos modelos (VOGADO et al., 2022). As imagens possuem o atributo Fotométrico com o parâmetro Monochrome2, onde pixels de maior intensidade representam cores claras e pixels de menor intensidade representam cores escuras. As Figuras 41 e 43 mostram exemplos de imagens das incidências frontais e laterais pertencentes do mesmo exame tanto para lombo-sacra quanto pododáctilos respectivamente. Cada incidência oferece informações específicas sobre a condição do paciente.

O conjunto de dados utilizado neste estudo abrange radiografias de lombo-sacra e pododáctilos, juntamente com seus relatórios médicos correspondentes, que foram produzidos por especialistas. Esses relatórios são divididos em seções importantes, como "DESCOBERTAS" e "IMPRESSÕES", que destacam as características clínicas relevantes observadas nas imagens e resumem as interpretações clínicas dos especialistas (PAVLOPOULOS et al., 2022). Nossa análise se concentra nessas seções para treinar os modelos generativos. As Figuras 41 e 43 apresentam exemplos de radiografias com seus respectivos relatórios médico, oferecendo uma visão completa dos dados.

Para desenvolver um modelo generativo eficaz, é fundamental entender as características do conjunto de dados. A Figura 29 apresenta a distribuição dos tamanhos dos laudos médicos de lombo-sacra, revelando um vocabulário de 1.881 palavras únicas, com um laudo contendo no máximo 1.026 palavras e no mínimo 15. A média de palavras por laudo é de 199.87, com desvio padrão de 58,70, fornecendo parâmetros importantes para a configuração do modelo generativo.

De forma semelhante, a Figura 30 mostra a distribuição dos laudos de pododáctilos, essencial para determinar o tamanho máximo de laudos que o modelo deve gerar. O conjunto de dados inclui 2.256 palavras únicas, com um laudo contendo até 1.167 palavras e um mínimo de 14, com média de 193,53 palavras e desvio padrão de 103,53, como exibido na figura. Essas informações são cruciais para configurar adequadamente o modelo gerador de laudos médicos.

Além da análise quantitativa dos laudos, é essencial explorar a frequência das palavras no conjunto de dados. As Figuras 31a e 31b mostram as palavras mais e menos frequentes, respectivamente. Palavras como "vértebras", "corpos" e "lombar" aparecem frequentemente, facilitando o aprendizado pelo modelo, enquanto termos raros como "descartar" e "flebólitos" apresentam maior desafio para a modelagem. A Figura 5.1.1 exibe uma nuvem de palavras com as 150 mais comuns, permitindo uma visualização

(a)

(b)

Laudo Médico
Mínimo desvio do eixo lombar para esquerda (provavelmente postural/posicional).
Corpos vertebrais de forma e contornos anatômicos.
Pedículos e apófises transversas sem alterações.
Discos intervertebrais conservados.

(c)

Medical Report
Minimum deviation of the lumbar axis to the left (likely postural/positional).
Vertebral bodies with anatomical shape and contours.
Pedicles and transverse processes without alterations.
Intervertebral discs preserved.

(d)

Figura 27 – Exemplo de um exame lombo-sacral com seu relatório médico: a) Visão frontal; b) Visão lateral; c) Relatório médico no idioma original (PT-BR); e d) Relatório médico traduzido para o inglês.

intuitiva da frequência das palavras.

De forma semelhante, os gráficos das Figuras 32a e 32b mostram as palavras mais e menos frequentes no conjunto de dados, com termos como “articulares”, “espaços” e “alterações” sendo comuns e mais fáceis para o modelo aprender. Já palavras raras como “lysfranc” e “calcaneocuboide”, que aparecem apenas uma vez, representam maior desafio na modelagem.

Para aprofundar nossa compreensão das palavras mais frequentes, criamos uma nuvem de palavras que destaca visualmente as 150 palavras mais comuns nos conjuntos de dados, conforme apresentado nas Figura 5.1.1 e 33.

5.1.2 Triagem das Imagens de Lombo-Sacra

Durante a aquisição do conjunto de dados de lombo-sacra, conforme ilustrado na Figura 23, enfrentamos vários desafios, incluindo variações na coloração dos ossos e do fundo das imagens, qualidade técnica limitada das radiografias e a presença de exames não pertinentes ao escopo deste estudo. Essas dificuldades originaram-se da diversidade das fontes de dados utilizadas para compilar o conjunto de dados. Todo esse processo é descrito na Seção 4.1.2

(a) (b) (c) (d)

Laudo Médico

COMENTÁRIOS
Ossos de textura e densidade normais.
Extremidades ósseas e interlinhas articulares preservadas.
Partes moles inalteradas.
Ausência de sinais de fraturas nas incidências obtidas.

IMPRESSÃO:
Estudo radiológico dos ossos dentro da normalidade.

RX PÉ ESQUERDO

COMENTÁRIOS
Ossos de textura e densidade normais.
Extremidades ósseas e interlinhas articulares preservadas.
Partes moles inalteradas.
Ausência de sinais de fraturas nas incidências obtidas.

IMPRESSÃO:
Estudo radiológico dos ossos dentro da normalidade.

(e)

Medical Report

COMMENTS
Bones of normal texture and density.
Bone ends and joint spaces are preserved.
Soft tissues are unaltered.
No signs of fractures in the obtained views.

IMPRESSION:
Radiological study of the bones within normal limits.

LEFT FOOT X-RAY

COMMENTS
Bones of normal texture and density.
Bone ends and joint spaces are preserved.
Soft tissues are unaltered.
No signs of fractures in the obtained views.

IMPRESSION:
Radiological study of the bones within normal limits.

(f)

Figura 28 – Exemplo de um exame de pododáctilos com seu laudo médico: a) Vista frontal esquerda; b) Vista frontal direita; c) Vista lateral direita; d) Vista lateral esquerda; e) Laudo médico na língua original português (PT-BR); e f) Laudo médico traduzido para inglês.

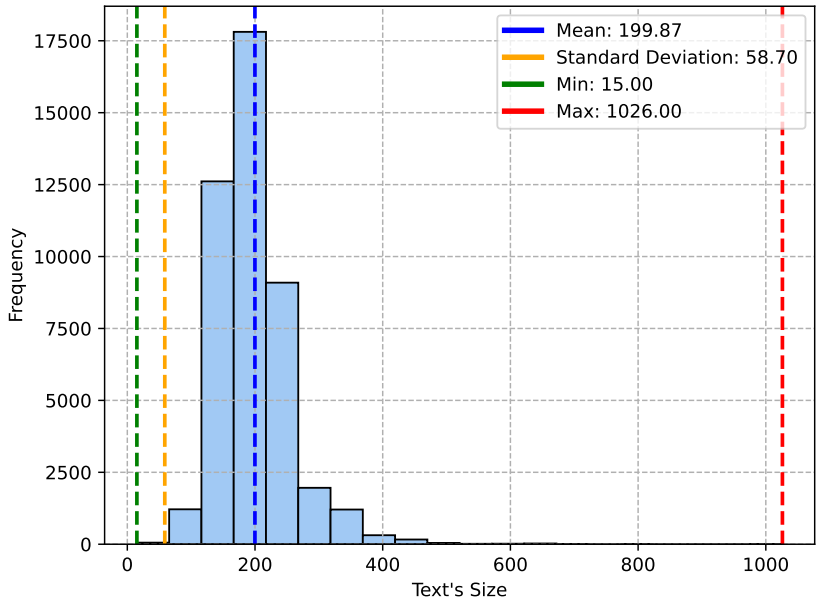


Figura 29 – Distribuição do texto no conjunto de dados de lombo-sacra.

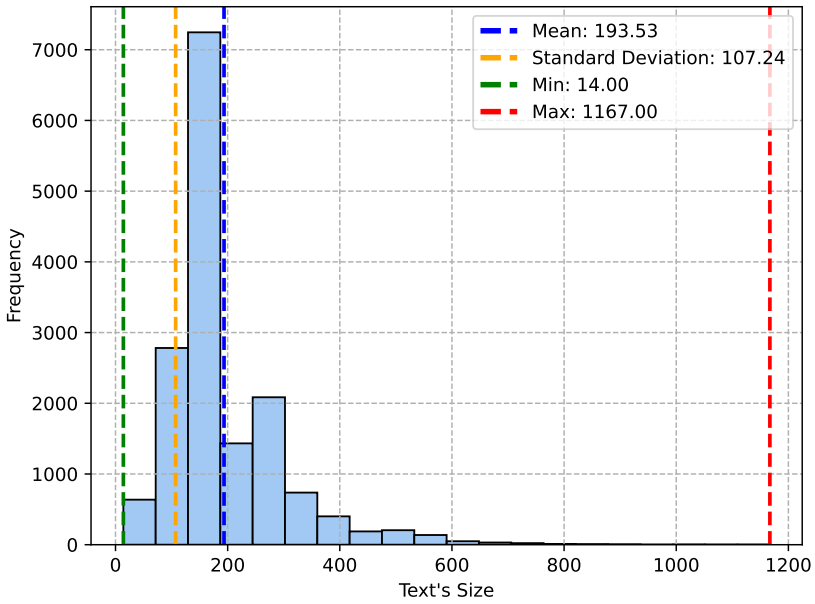
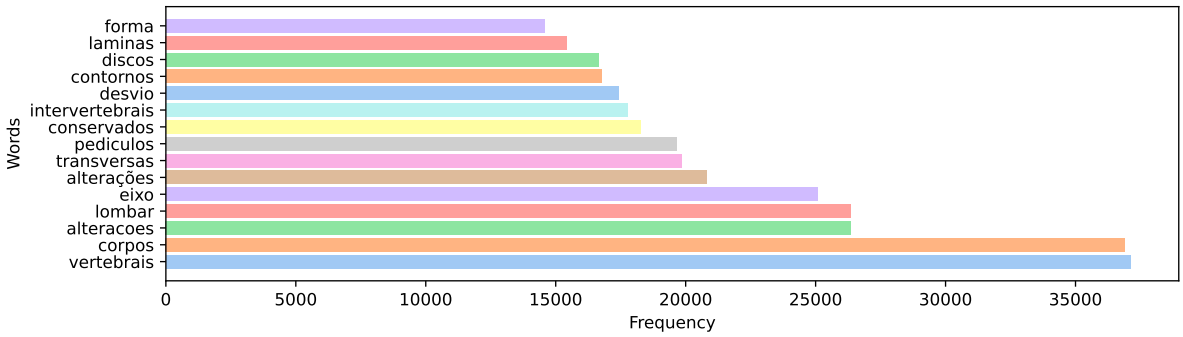
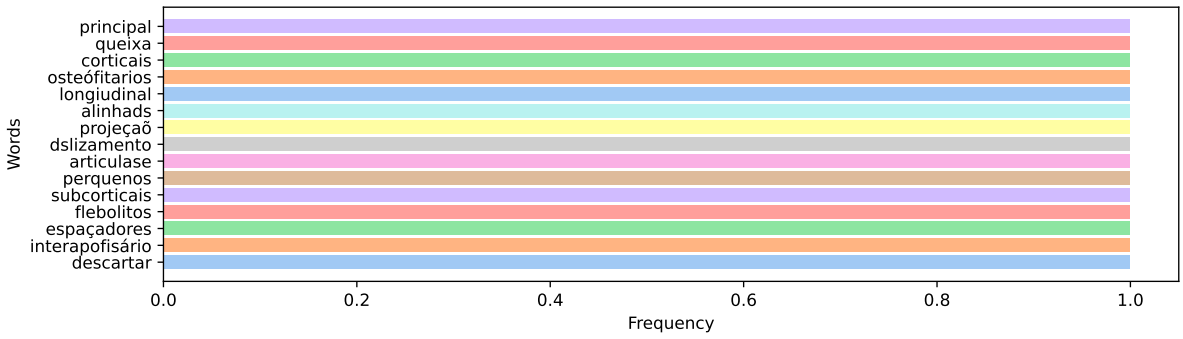


Figura 30 – Distribuição do texto no conjunto de dados de pododáctilos.



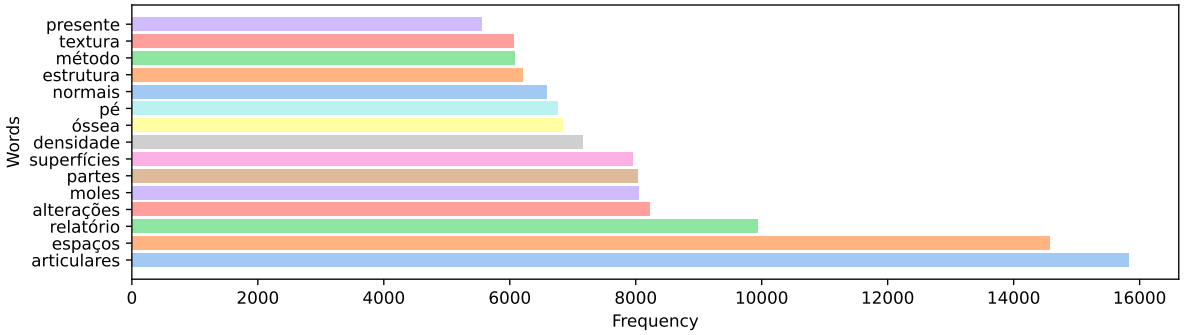
(a)



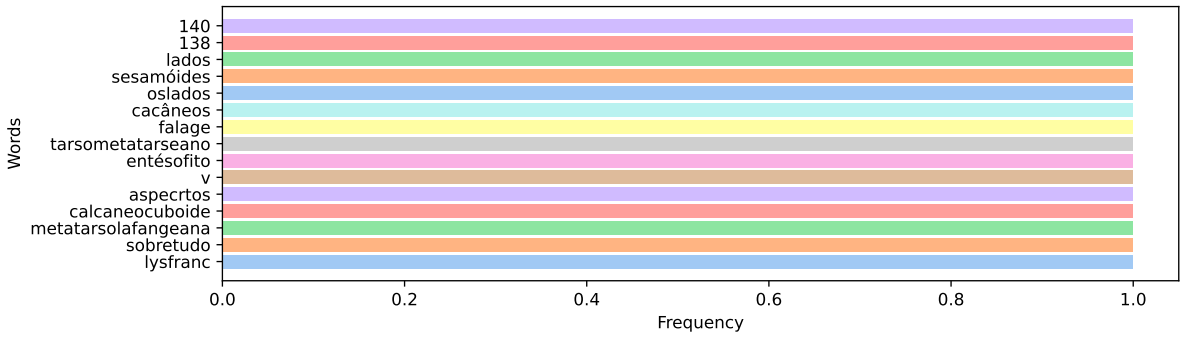
(b)

Figura 31 – Gráfico das palavras mais e menos frequentes no conjunto de dados de lombo-sacra.

Na primeira etapa, para lidar com as variações de coloração dos ossos, seguimos a metodologia descrita por (VIEIRA et al., 2023a), que consistiu no treinamento de uma CNN pré-treinada, VGG16, em combinação com um classificador MLP. Essa abordagem



(a)



(b)

Figura 32 – Gráfico das palavras mais e menos frequentes no conjunto de dados de pododáctilos.

de palavras representando as 150 palavras mais frequentes no conjunto de dados de lombo-sacra.

foi desenvolvida para padronizar a coloração dos ossos para branco e o fundo para preto, garantindo conformidade com o padrão utilizado pelos especialistas na interpretação das radiografias.

Em seguida, foi realizado a classificação de incidências para garantir a triagem das radiografias da coluna lombo-sacra e a identificação da orientação (frontal ou lateral) das imagens, empregamos o método também descrito por Vieira et al. (2023a). Este

Figura 33 – Nuvem de palavras representando as 150 palavras mais frequentes no conjunto de dados de pododáctilos.

método utilizou a ResNet50 pré-treinada em conjunto com um classificador de SVM.

5.1.3 Pré-processamento de Imagens

Para otimizar o processo de geração automática de laudos médicos preliminares em radiografias de lombo-sacra e pododáctilos, conforme ilustrado nas Figuras 34 e 35, implementamos uma série de técnicas de pré-processamento.

O conjunto de dados utilizado neste estudo provém de múltiplas fontes, sendo gerado por diferentes equipamentos e técnicos, resultando em imagens com características heterogêneas, como variações em tamanho, contraste, brilho e enquadramento. Para melhorar o aprendizado do modelo proposto, realizamos uma etapa de pré-processamento visando padronizar essas características. As Figuras 34 e 35 demonstram os efeitos desse pré-processamento, evidenciando como as imagens brutas foram transformadas em um formato mais uniforme e adequado para análise, aumentando assim a confiabilidade e a robustez da metodologia.

Adotamos aqui a mesma metodologia de pré-processamento aplicada no método de detecção de anomalias em lombo-sacra, conforme descrito na Seção 4.1.3. Essa replicação foi fundamental para garantir a consistência nos procedimentos e nos resultados, facilitando a comparação e a integração dos modelos em diferentes etapas do estudo.

5.1.4 Pré-processamento de Texto e Análise

Os laudos presentes neste conjunto de dados foram elaborados por diferentes especialistas, com variações em idade, nível de experiência, especialidade e condições de trabalho, o que resulta em uma diversidade textual significativa. Essa variação pode gerar textos distintos para casos semelhantes. Para otimizar o aprendizado do modelo

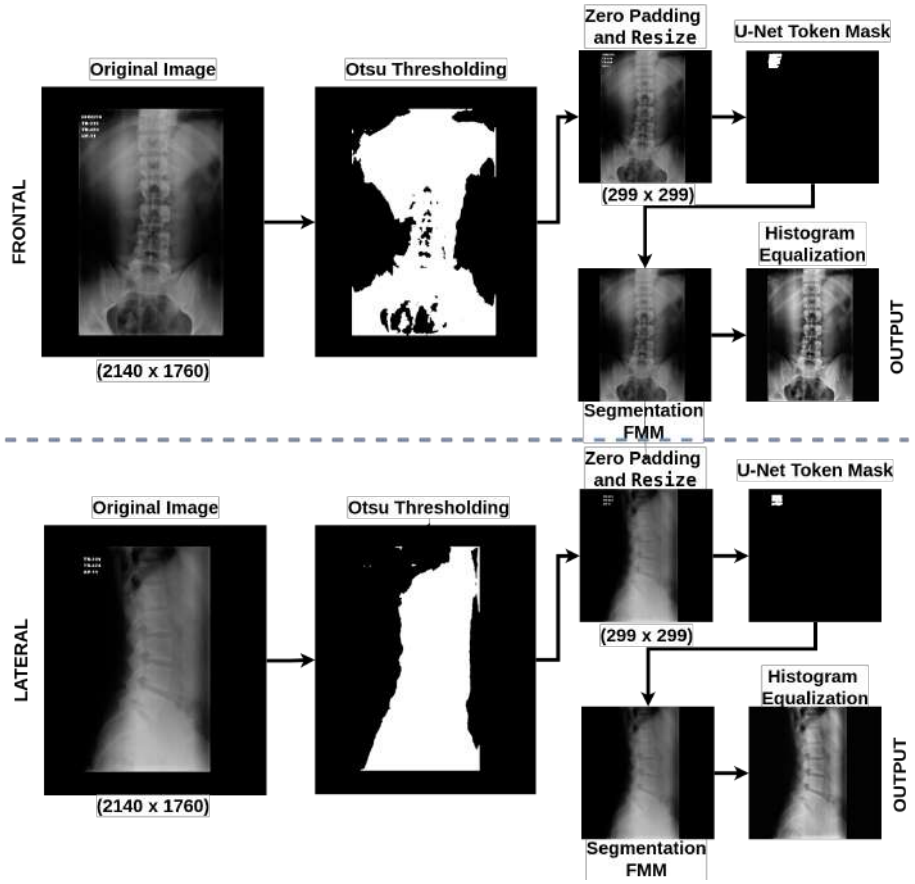


Figura 34 – Fluxo de pré-processamento para exame radiográfico de lombo-sacra nas incidências frontal e lateral. Inicialmente, foi empregado o limiarização de Otsu para eliminar bordas externas. Posteriormente, foi aplicado um preenchimento de zeros para alcançar um redimensionamento não distorcido em formato quadrado. Em seguida, os marcadores metálicos foram segmentados utilizando uma arquitetura U-Net, as regiões segmentadas foram preenchidas usando o FMM. E finalmente, a imagem resultante foi submetida à equalização de histograma com o CLAHE. Fonte da Figura: (VIEIRA et al., 2023a).

proposto, também implementamos uma etapa de pré-processamento específica para os laudos médicos.

Na primeira etapa, referente à remoção de informações irrelevantes, eliminamos elementos desnecessários do texto, como carimbos frequentemente adicionados ao final de relatórios padronizados por profissionais, clínicas ou hospitais. Um exemplo comum é "*OBS.: Exame documentado em CD*", entre outros.

Na próxima etapa, relativa à adição de marcadores, incluímos indicadores que informam ao modelo o início e o fim de um relatório, com os marcadores [*START*] para iniciar e [*END*] para encerrar, definindo os pontos exatos para a geração do relatório.

Na etapa posterior, correspondente à análise quantitativa dos laudos, avaliamos os relatórios para extrair informações relevantes que auxiliam na definição de hiperparâmetros do modelo. Por exemplo, observamos dados quantitativos sobre a estrutura das palavras

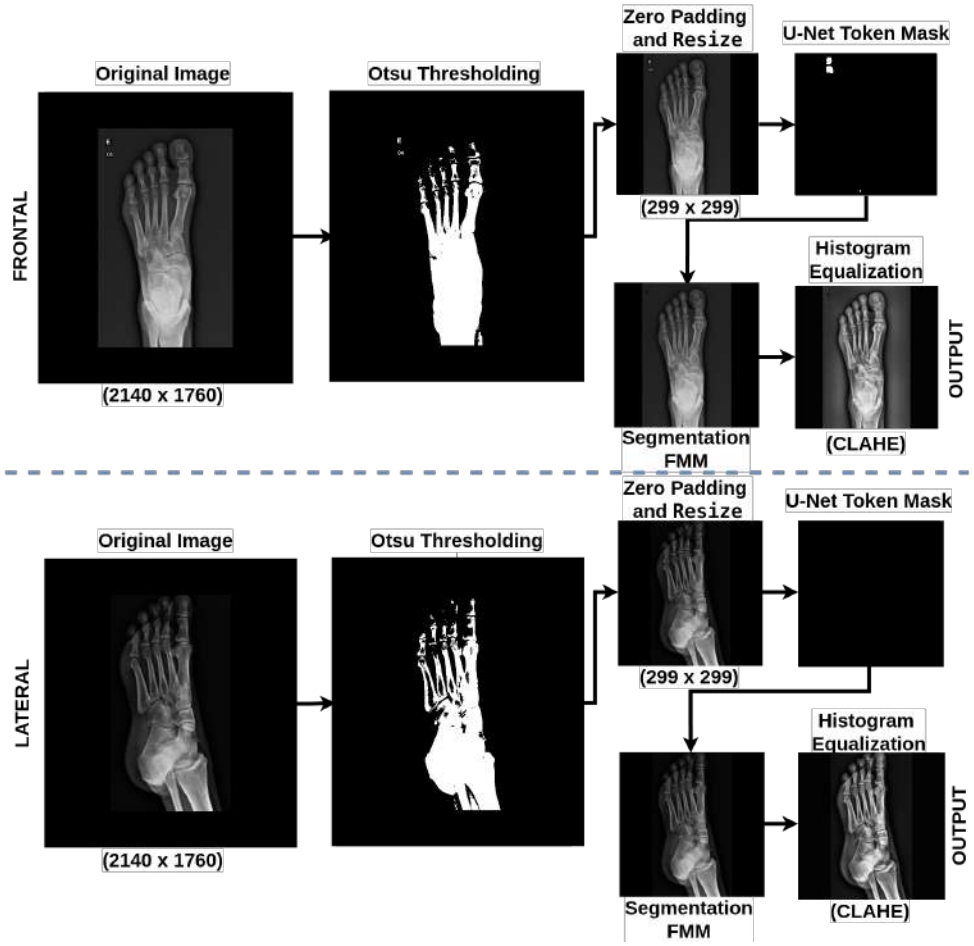


Figura 35 – Fluxo de pré-processamento para exame radiográfico de polidactilias nas incidências frontal e lateral. Inicialmente, foi empregado o limiarização de Otsu para eliminar bordas externas. Posteriormente, foi aplicado um preenchimento de zeros para alcançar um redimensionamento não distorcido em formato quadrado. Em seguida, os marcadores metálicos foram segmentados utilizando uma arquitetura U-Net, as regiões segmentadas foram preenchidas usando o FMM. E finalmente, a imagem resultante foi submetida à equalização de histograma com o CLAHE.

no conjunto de dados, conforme apresentado na Tabela 12.

Tabela 12 – Informações quantitativas sobre os laudos de lombo e pododáctilos.

Informação	**Lombo-Sacra**	**Pododáctilos**
Palavras únicas	1,883	2,256
Maior número de palavras em um laudo	1,026	1,167
Menor número de palavras em um laudo	15	14
Média de palavras por laudo	199,87	193,53
Desvio padrão da contagem de palavras	58,70	107,24

Em seguida, na definição do vocabulário, com base nos dados da Tabela 12, determinamos o tamanho do vocabulário a ser usado pelo modelo em seu repertório, através da função `choose_vocabulary_size` 5.1.

$$\text{choose\_vocabulary\_size}(df, \text{column\_name}) = \min\left(\left|\text{set}(\text{words})\right|, |df|\right), \tag{5.1}$$

onde, df é o vetor contendo os textos, column_name é o nome da coluna contendo os textos, words representa as palavras extraídas dos textos na coluna especificada de df, set(words) refere-se ao conjunto de palavras únicas, $|\text{set}(\text{words})|$ corresponde ao tamanho desse vocabulário único, e $|df|$ é o número de amostras no conjunto de dados, utilizado para limitar o tamanho do vocabulário, se necessário.

Definição do Comprimento Máximo de Sentença. Além disso, os dados apresentados na Tabela 12 permitiram estabelecer o comprimento máximo de uma sentença utilizando a Fórmula para `max_length` 5.2.

$$\text{max\_length} = \text{int}\left(\frac{\text{max\_dim} + \text{min\_dim}}{2}\right), \tag{5.2}$$

onde, max_dim é o tamanho máximo de palavras em um laudo, min_dim é o tamanho mínimo de palavras em um laudo.

Definição das Dimensões do Embedding. Os dados fornecidos pela Tabela 12 também nos permitiram definir as dimensões do embedding por meio da Função `choose_embedding_dim` 5.3, que pode ser representada matematicamente como,

$$\begin{aligned}&\text{choose\_embedding\_dim}(df, \text{column\_name}, \text{min\_dim}, \text{max\_dim})\\ &\quad= \text{int}\left((\text{max\_dim} - \text{min\_dim}) \times \frac{\text{total\_words}}{\text{len}(df)} + \text{min\_dim}\right),\end{aligned} \tag{5.3}$$

onde, df é o vetor contendo os textos, column_name é o nome da coluna contendo os textos, min_dim é o tamanho mínimo de palavras em um laudo, max_dim é o tamanho máximo de palavras em um laudo, total_words é o número total de palavras únicas em todo o conjunto de dados, $\text{len}(df)$ é o número de amostras no conjunto de dados.

Finalizamos o processamento do texto aplicando *Embedding*, tokenização e *shifted sequence* nos alvos. O *Embedding* transforma palavras em vetores densos, capturando seu significado semântico (BUTNARU; IONESCU, 2017; GONG; COSMA; FANG, 2021). A tokenização, converte o texto em sequências de *tokens*, simplificando o processamento pelo modelo e permitindo a geração de textos correspondentes às imagens de entrada em modelos generativos *image-to-text* (PAVLOPOULOS et al., 2022; ISLAM et al., 2024). Para tornar o processo de aprendizado mais robusto, modelos de *Transformers* geralmente empregam o "*shifted sequence*" nos textos alvos. Essa técnica

desloca as sequências de texto durante o treinamento, incentivando o modelo a prever o próximo *token* em uma sequência deslocada, o que melhora sua capacidade de generalização e precisão na geração de texto.

5.1.5 Aumento de Dados

Embora o conjunto de dados utilizado neste estudo possua uma quantidade maior que a encontrada no estado-da-arte, para a aplicação envolvendo modelos generativos em radiografias lombo-sacras e pododáctilos, nossa metodologia ainda requer uma ampla variedade de amostras devido à complexidade do método adotado (PAVLOPOULOS et al., 2022). Portanto, a utilização de técnicas de aumento de dados é essencial para aprimorar o aprendizado do modelo proposto (CONNOR, 2019; VIEIRA et al., 2023a). Nas Figuras 36 e 37 , são apresentadas as técnicas de aumento de dados, juntamente com exemplos de seus resultados, adotadas neste estudo para a geração de imagens sintéticas com base em imagens reais.

Nossa metodologia incorpora diversas técnicas de aumento de dados, com parâmetros específicos projetados para assegurar que as amostras sintéticas sejam representações plausíveis de cenários do mundo real. Estas técnicas simulam variações em enquadramentos, posições e qualidade das imagens, preservando as características essenciais do exame. As técnicas foram aplicadas tanto para as imagens frontais quanto laterais dos exames e incluem: escala negativa de -10% e positiva de +10%; Rotação de 15% para a direita e esquerda; Ruído gaussiano de no máximo 10%; Deslocamento de 4%; e equalização de histograma.

Todas estas técnicas de aumento de dados foram aplicadas usando o “ImageDataGenerator” do TensorFlow, que atribui uma probabilidade de 50% para que cada operação de aumento seja aplicada ou não em cada imagem. Quando uma técnica é selecionada, os valores de seus parâmetros são definidos aleatoriamente dentro do intervalo previamente estabelecido. Dessa forma, garantimos que as imagens sintéticas sejam suficientemente variadas, sem que nenhuma técnica seja necessariamente repetida em todas as amostras. As Figuras 37 e 37 ilustram exemplos deste processo.

Durante os treinamentos do modelo, utilizamos aumento de dados em tempo de execução, onde as operações matemáticas são aplicadas aleatoriamente nas imagens do conjunto de treinamento para criar variações sintéticas a cada época. As imagens originais não são apresentadas ao modelo durante o treinamento, sendo utilizadas apenas nos conjuntos de teste e validação. Assim, garantimos que o modelo aprenda a partir de diferentes variações, o que pode aumenta a robustez e a capacidade de generalização do modelo.

Para garantir a heterogeneidade do conjunto de dados e melhorar a capacidade

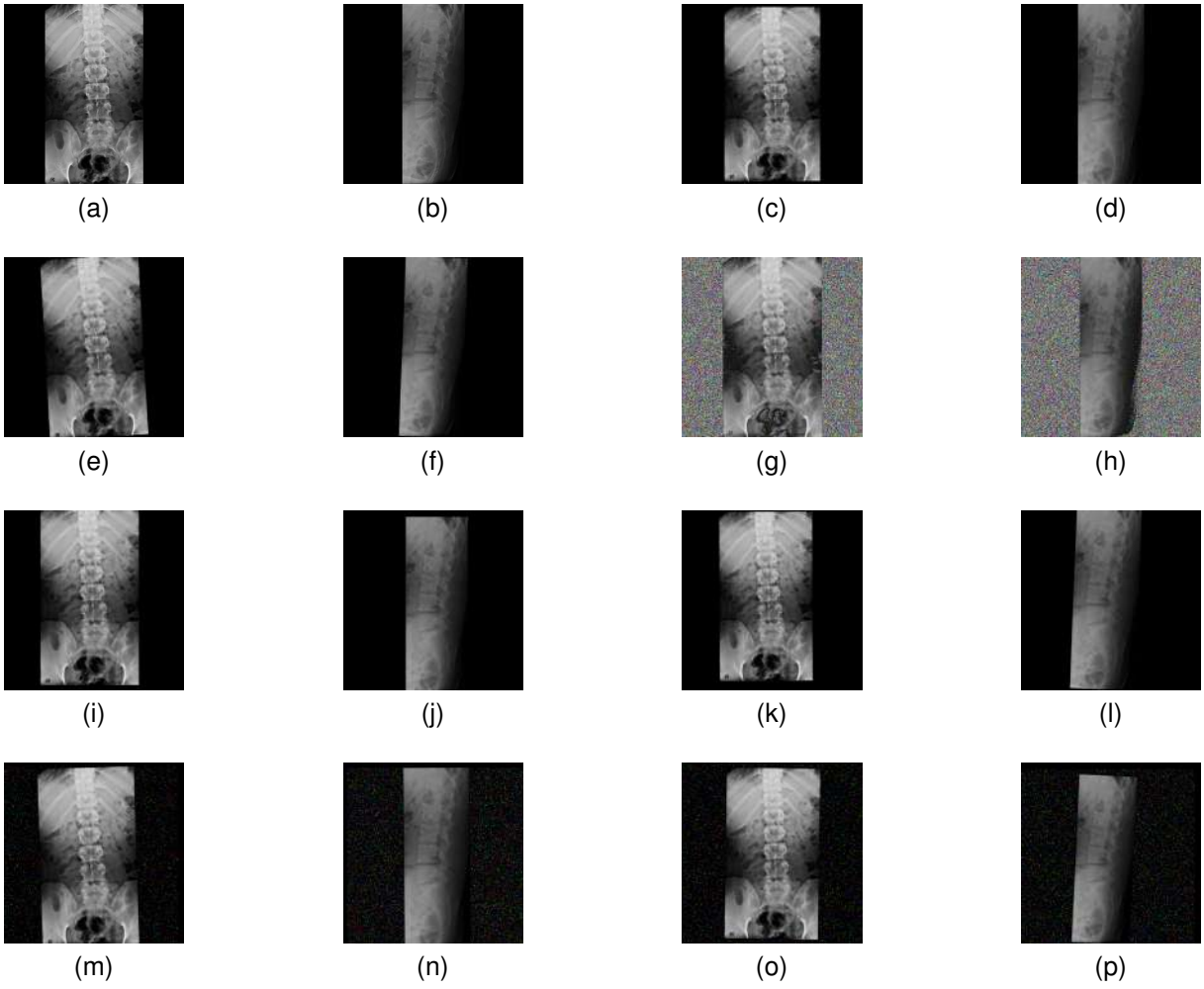
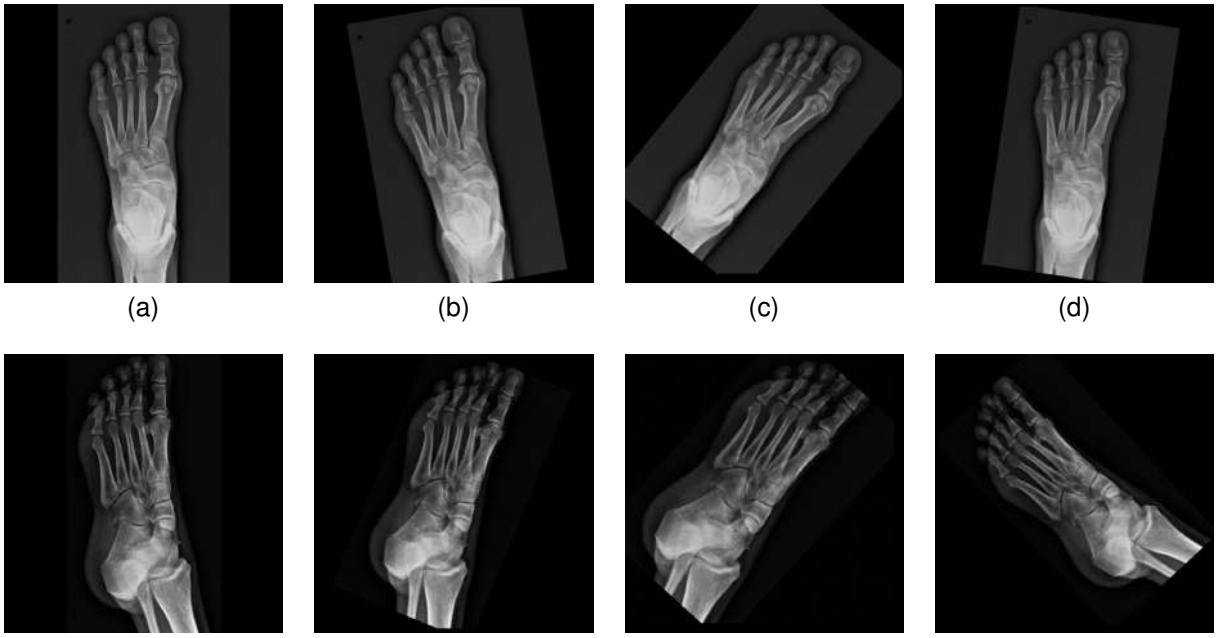


Figura 36 – Exemplos de aumento de dados frontal e lateral: (a, b) Imagens originais; (c, d) Escala; (e, f) Rotação; (g, h) Ruído gaussiano; (i, j) Deslocamento; (k - p) Combinação aleatória de todas as operações.

do modelo de lidar com diferentes variações, utilizamos técnicas de aumento de dados conforme descrito. No entanto, é importante ressaltar que o objetivo principal dessas técnicas não foi mitigar o desbalanceamento dos dados, mas sim tornar o conjunto de dados mais diversificado e representativo de cenários clínicos reais. No caso das radiografias de lombo-sacra, enfrentamos um desbalanceamento com 39,9% de casos normais e 60,1% de casos anormais, enquanto no conjunto de dados de pododáctilos, o desbalanceamento é ainda mais acentuado, com 44,2% de exames normais e 55,8% de exames anormais. Acreditamos que essa abordagem, ao fornecer um conjunto de dados mais rico e variado, auxilia o modelo a lidar com o desbalanceamento inerente, promovendo um aprendizado generalizável para a detecção de anomalias e características clínicas distintas em ambos os contextos.



(a) (b) (c) (d)

(e) (f) (g) (h)

Figura 37 – Exemplo de aumento de dados de polidactilia: (a, e) Imagens frontal e lateral originais; (b-d) exemplos de amostras sintéticas geradas a partir de (a) usando técnicas aleatórias de aumento de dados; (f-h) exemplos de amostras sintéticas geradas a partir de (b) usando técnicas aleatórias de aumento de dados.

5.1.6 Rede Neural Convolucional Aplicada

Neste estudo, optamos por utilizar a arquitetura Inception-V3 para a extração de características das imagens, tanto nas incidências frontais quanto laterais. Essa escolha foi motivada pela disponibilidade de pesos pré-treinados no conjunto de dados ImageNet (KRIZHEVSKY; SUTSKEVER; HINTON, 2017), aproveitando o conhecimento adquirido a partir de milhões de imagens. A Inception-V3 é uma rede neural profunda que utiliza convoluções e módulos Inception para capturar informações em várias escalas e níveis de abstração, permitindo o aprendizado de representações complexas de objetos e características nas imagens (SZEGEDY et al., 2015). A arquitetura possui um total de 21.802.784 parâmetros, dos quais 21.768.352 são treináveis, e 34.432 são não treináveis.

Embora a VGG16 tenha obtido o melhor desempenho na triagem de anomalias em lombo-sacra, optamos por não utilizá-la devido ao tamanho considerável de seus arquivos de pesos, o que inviabilizou seu uso em nosso hardware e no pipeline planejado. A VGG16 requer uma grande quantidade de recursos computacionais, o que estava fora do escopo dos nossos recursos disponíveis. Dessa forma, empregamos a Inception-V3, que apresentou um bom compromisso entre desempenho e viabilidade computacional, garantindo que o processo de extração de características fosse eficiente e adequado ao nosso fluxo de trabalho.

Além disso, a capacidade do Inception-V3 de extrair representações de alto nível é essencial para tarefas como a conversão de imagem em texto, alinhando-se diretamente ao nosso objetivo de automatizar a geração de laudos médicos para radiografias de pododáctilos.

No processo de treinamento, adotamos uma abordagem híbrida. Inicialmente, seguimos uma metodologia de treinamento rasa, liberando, na segunda etapa, todas as camadas convolucionais das CNNs para que pudesse aprender com o novo problema proposto. Dessa forma, as CNNs apenas atualizaram seus *kernels* após o modelo ter absorvido o conhecimento do novo conjunto de dados, utilizando a segunda etapa para aprofundar o processo de aprendizagem.

Na primeira etapa de treinamento, realizamos o ajuste fino dos *Transformers* enquanto mantínhamos as camadas convolucionais (CNNs) treinandas no ImageNet congeladas, de modo que apenas os *Transformers* aprendessem as características do conjunto de dados. Para essa fase, utilizamos um valor de alpha [1] de 0,001, com um decaimento [2] de 0,1, para garantir uma convergência rápida e eficaz. A técnica de *early stopping* foi empregada para evitar o sobreajuste, interrompendo o treinamento após detectar o início da deterioração no desempenho do conjunto de validação, possibilitando uma execução mínima de 8 épocas. Optamos pelo otimizador ADAM (KINGMA; BA, 2017), devido à sua capacidade de ajuste automático das taxas de aprendizado, enquanto monitorávamos as métricas de acurácia e perda por *Sparse Categorical Crossentropy*. Nesta etapa, o tamanho de lote foi de 16, com um *buffer* de 4.300 para garantir a aleatorização das amostras.

Na segunda etapa de treinamento, descongelamos as camadas convolucionais das CNNs para realizar um ajuste fino profundo, permitindo que tanto os *Transformers* quanto as CNNs aprendessem características mais específicas e sutis do conjunto de dados. Nesta etapa, reduzimos o alpha para 0,0001, mantendo o mesmo decaimento de 0,1, de forma a ajustar os pesos com maior precisão e capturar detalhes que não foram apreendidos na fase anterior. O objetivo dessa abordagem foi aumentar a capacidade de aprendizado do modelo, utilizando o ajuste fino tanto das representações visuais quanto das linguísticas.

Durante ambas as etapas de treinamento, adotamos o otimizador ADAM para melhorar a convergência dos modelos de forma eficiente. A técnica de *early stopping* foi essencial para mitigar o risco de sobreajuste, interrompendo o treinamento no momento certo. Estabelecemos um tamanho de lote de 16 e um *buffer* de 4.300,

[1] Parâmetro que controla a magnitude dos ajustes nos pesos do modelo durante a otimização, influenciando a taxa de convergência no treinamento.

[2] O decaimento do alpha se refere à redução gradual da taxa de aprendizagem durante o treinamento do modelo, ajustando-a para valores menores ao longo do tempo para melhorar a estabilidade e eficiência da convergência.

configuração determinada pela limitação dos 16 GB de RAM da GPU, que não permitia o processamento simultâneo de lotes maiores contendo o modelo, as imagens e os textos.

5.1.7 Processo de Transição Imagem-Texto

O processo de transição das características visuais extraídas de uma imagem para a geração de texto no modelo envolve a combinação entre arquiteturas de CNNs e *Transformers*, que trabalham de maneira integrada para converter informações visuais em descrições textuais coerentes e detalhadas (VASWANI et al., 2017). Esse processo pode ser dividido em várias etapas principais, conforme descrito a seguir.

Na primeira etapa, as características visuais são extraídas das imagens por meio de CNNs pré-treinadas, que capturam as representações intrínsecas e relevantes das imagens durante o treinamento (KRIZHEVSKY; SUTSKEVER; HINTON, 2017). A imagem passa por uma série de camadas convolucionais, que aprendem a detectar padrões hierárquicos, como bordas, texturas e estruturas complexas. As saídas dessas camadas convolucionais são então convertidas em vetores de características por meio da operação de *Flatten*, resultando em uma representação densa e compacta para cada imagem. Para radiografias de lombo-sacra, que possuem duas incidências (frontal e lateral) e utilizam duas CNNs Inception-V3, o vetor de características tem tamanho 262.144. Já para radiografias de pododáctilos, que possuem quatro imagens (frontal e lateral para os pés direito e esquerdo) e utilizam quatro CNNs Inception-V3, o vetor resultante possui tamanho 524.288 (VIEIRA et al., 2021).

Em seguida, esses vetores formam uma representação unificada que encapsula a informação visual relevante das imagens analisadas. Essa representação é, então, processada por um codificador baseado em uma arquitetura de *Transformer*, que utiliza mecanismos de normalização em camadas (*Layer Normalization*) e atenção multi-cabeça (*multi-head attention*) para refinar ainda mais as características visuais. O objetivo do codificador é preservar a integridade das informações extraídas, enquanto captura as dependências contextuais entre diferentes partes da imagem, garantindo uma compreensão mais ampla das relações entre as características visuais (MOHSAN et al., 2023; VASWANI et al., 2017).

Após o processo de codificação, o vetor resultante é direcionado para um decodificador baseado em *Transformer*, que começa a geração da descrição textual de forma sequencial, palavra por palavra. Utilizando uma máscara causal de atenção (*causal attention mask*), o decodificador garante que a geração seja autorregressiva, ou seja, que cada palavra gerada dependa das palavras anteriores na sequência. O decodificador incorpora *embeddings* de palavras e aplica múltiplas camadas de atenção para integrar as informações visuais do codificador com as palavras geradas anteriormente. Essas integrações são transformadas em probabilidades sobre o vocabulário de saída por meio

de camadas densas *feed-forward* (TSANIYA; FATICHAH; SUCIATI, 2024; CAO et al., 2023; VASWANI et al., 2017).

Durante o treinamento, a entrada do decodificador inclui a sequência de palavras até a palavra atual, e o modelo é treinado para prever a próxima palavra. A função de perda é calculada comparando as palavras geradas com a sequência alvo, ajustada pela máscara de atenção. Durante a inferência, o modelo utiliza cada palavra gerada como entrada para o próximo passo, continuando a geração até encontrar o *token* que marca o fim da sequência (VASWANI et al., 2017).

Esse mecanismo integrado permite que o modelo capture de maneira eficaz a transição das características visuais para descrições textuais, garantindo que os laudos médicos gerados sejam precisos, ricos em contexto e semanticamente coerentes.

5.1.8 Desenho dos Experimentos

Considerando os desafios computacionais enfrentados pela metodologia desenvolvida neste estudo, que visa desenvolver e testar modelos generativos para gerar automaticamente laudos médicos a partir de radiografias de lombo-sacra e pododáctilos, projetamos experimentos que abordam esses desafios. Essa abordagem garante que a estrutura experimental avalie rigorosamente o desempenho do modelo na geração de laudos médicos precisos e confiáveis, além de otimizar a eficiência computacional do sistema.

Para a avaliação do desempenho de nosso método tanto para radiografias de lombo-sacra quanto de pododáctilos, adotamos a metodologia *K-Fold* com diferentes configurações de *folds*. No caso de lombo-sacra, utilizamos 5 *folds*. As amostras foram aleatoriamente embaralhadas e divididas em 5 conjuntos de tamanhos semelhantes. Em cada iteração, um conjunto foi retido para validação e os demais usados para treinamento, repetindo o processo 5 vezes, de modo que cada conjunto foi utilizado uma vez para validação. Já para as radiografias de pododáctilos, a metodologia *K-Fold* foi aplicada com 10 *folds*, onde as amostras foram divididas em 10 conjuntos de tamanhos semelhantes, repetindo o processo 10 vezes. Em ambas as abordagens, os resultados das iterações foram agregados para avaliar a arquitetura, utilizando o desempenho médio como índice de avaliação do método (KOHA, 1995; SARAIVA1 et al., 2019). Embora computacionalmente mais custosa, essa estratégia maximiza o uso dos dados e permite uma melhor avaliação da capacidade de generalização do modelo, fornecendo uma medida mais robusta e confiável do desempenho.

Utilizamos diferentes configurações de *folds* para cada tipo de exame devido ao custo computacional envolvido em cada conjunto de dados. Para as radiografias de lombo-sacra, que possuíam uma maior quantidade de amostras, optamos por 5 *folds* para

equilibrar a carga de processamento. Já para as radiografias de pododáctilos, que apresentavam um número menor de amostras, utilizamos 10 *folds* para maximizar a avaliação do modelo. O maior limitador não foi a GPU, mas sim a biblioteca de cálculo das métricas BLEU, METEOR e ROUGE, aac-metrics, que demandava um uso computacional extremo do processador, podendo até levar a complicações no hardware.

Os dados dos experimentos, tanto para lombo-sacra quanto para pododáctilos, foram divididos em conjuntos de treinamento, validação e teste. Para lombo-sacra, a divisão foi de 70%, 10% e 20%, respectivamente, enquanto para pododáctilos, os conjuntos foram divididos em 80%, 10% e 10%. Em ambos os casos, o conjunto de teste foi reservado para avaliar a capacidade de generalização do modelo para novos casos. Em cada *fold*, o *Transformer* foi treinado por até 20 épocas. Na primeira etapa, o *Transformer* aprendeu as características do problema, enquanto as camadas convolucionais das CNNs permaneceram congeladas. Após essa etapa, as CNNs foram descongeladas e treinadas por até 10 épocas adicionais, permitindo que também aprendessem características do conjunto de dados.

Para este trabalho, utilizamos o ambiente computacional com as seguintes configurações: 64 GB de RAM; CPU: 12 núcleos; GPU: RTX 4080, 2505 MHz, 9728 CUDA Cores, 16 GB 22.4 Gbps; SO: Linux Ubunto 23.10; Linguagem de programação: Python 3.10.12; Bibliotecas: Tensorflow 2.15.0, Keras 2.15.0, Opencv 4.8.1, Sklearn 1.3.2, Numpy 1.26.2, Skimage 0.22.0, PILL 10.1.0, TQDM 4.66.1, NLTK 3.8.1, aac-metrics 0.5.4.

5.2 Resultados Para Geração Automática de Laudos Preliminares em Lombo-Sacra

Nesta seção, apresentamos os resultados obtidos pela metodologia para laudos médicos em radiografias lombo-sacrais, conforme mostrado na Tabela 13. Em seguida, comparamos nossos resultados com o estado-da-arte na Tabela 14.

Vamos analisar os resultados de nossa metodologia com base nas métricas de validação da Tabela 13, que apresenta a média e o desvio padrão de cada métrica em 5 *folds*, para as classes “Anormal” e “Normal” separadamente e para o conjunto de ambas. Essa análise mais detalhada por classe oferece uma compreensão completa do desempenho do modelo em diferentes cenários clínicos, fornecendo informações valiosas para ajustes futuros e para orientar decisões clínicas específicas na aplicação do sistema de geração automática de laudos radiológicos preliminares.

Ao analisar os resultados da classe “Anormal”, as pontuações das métricas BLEU-1 a BLEU-4, embora ligeiramente menores que as da classe “Normal”, permanecem acima de 0,4, o que indica uma correspondência razoável entre os laudos gerados pelo modelo e os laudos humanos. No entanto, as métricas METEOR e ROUGE-L são mais

Tabela 13 – Resultados médios das métricas para cada *fold*, juntamente com o desvio padrão.

LABEL	BLEU-1	BLEU-2	BLEU-3	BLEU-4	METEOR	ROUGE-L
ANORMAL	0,549±0,018	0,486±0,019	0,437±0,019	0,399±0,018	0,370±0,013	0,580±0,018
NORMAL	0,691±0,033	0,636±0,036	0,596±0,040	0,561±0,042	0,598±0,038	0,701±0,031
TODOS	0,612±0,018	0,552±0,020	0,507±0,021	0,470±0,022	0,471±0,019	0,633±0,018

baixas, sugerindo menor concordância linguística e similaridade textual, possivelmente devido à maior complexidade das anormalidades radiográficas e seus laudos, que apresentam padrões mais variados e desafiadores para a geração automática de laudos preliminares.

Em contraste com os resultados da classe “Anormal”, a classe “Normal” apresentou desempenho superior. As métricas BLEU-1 a BLEU-4 foram mais elevadas, sugerindo uma correspondência mais precisa entre os laudos gerados pelo modelo e os laudos humanos. Além disso, a métrica METEOR e o ROUGE-L também indicaram maior concordância linguística e textual, confirmando o que esperávamos que os laudos de exames normais tendem a ser mais uniformes, o que facilita uma correspondência mais robusta entre os laudos gerados e os laudos reais.

Nos resultados da combinação das classes, as métricas BLEU-1 a BLEU-3 superaram 0,5, enquanto o BLEU-4 obteve 0,470, indicando uma correspondência significativa entre os relatórios gerados e os humanos. A métrica METEOR atingiu 0,471, reforçando a concordância linguística, e o ROUGE-L alcançou 0,633, demonstrando alta similaridade entre as sequências de palavras dos relatórios gerados e humanos.

A qualidade dos resultados obtidos deve-se à integração de duas CNNs pré-treinadas com transformadores, permitindo ao modelo capturar nuances das imagens frontais e laterais e correlacioná-las com as informações visuais e linguísticas presentes nos laudos médicos.

A análise dos dados na Figura 38 revela uma distribuição consistente das métricas de avaliação (BLEU-1, BLEU-2, BLEU-3, BLEU-4, METEOR e ROUGE-L) ao longo das 5 dobras da validação cruzada, com medianas relativamente altas e variações ocasionais devido a *outliers*. No entanto, a estabilidade geral das métricas, incluindo METEOR e ROUGE-L, sugere que o modelo mantém um desempenho uniforme e confiável, estabelecendo uma base sólida para a robustez e confiança do modelo proposto na geração de relatórios médicos automáticos.

A análise do *box-plot* mostra que as classes “Anormal” e “Normal” possuem distribuições de métricas semelhantes ao longo das dobras, indicando um desempenho estável na geração de relatórios médicos. No entanto, a classe “Anormal” apresenta maior dispersão nos dados e intervalos interquartis mais amplos em comparação à classe “Normal”, sugerindo maior variabilidade nas previsões para radiografias anormais.

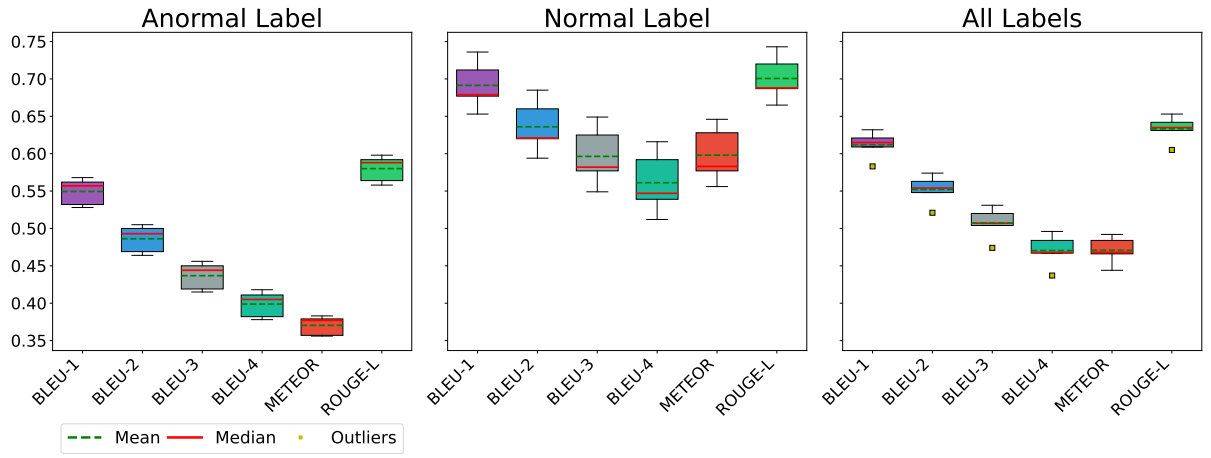


Figura 38 – *Box plot* dos resultados da geração automática de laudos médicos preliminares com as métricas Bleu-1, Bleu-2, Bleu-3, Bleu-4, Meteor e Rouge-L ao longo dos 5 *k-folds*, para as classes “Anormal” e “Normal”, bem como para a combinação de ambas.

Ao considerarmos a combinação de ambas as classes, a distribuição das métricas se alinha estreitamente com a das classes individuais. Os valores medianos permanecem acima de 0,45, indicando desempenho robusto em todas as dobras. A consistência na distribuição das métricas entre as classes “Anormal”, “Normal” e a combinação das duas classes reforça ainda mais a confiabilidade e estabilidade do modelo na geração de relatórios médicos automáticos em diferentes cenários clínicos.

A análise comparativa entre nossa abordagem proposta e os métodos estado-da-arte é detalhada na Tabela 14. Em várias métricas de avaliação, nosso método se destaca consistentemente, evidenciando sua eficácia na geração de relatórios médicos. No entanto, ao compararmos com o estudo de Cao et al. (2023), que explorou o uso de imagens de endoscopia, observamos resultados superiores em métricas específicas. Vale ressaltar que, ao contrário de nossa abordagem, o estudo de Cao et al. (2023). não utiliza a métrica METEOR em sua avaliação, o que pode limitar a compreensão abrangente de seu desempenho. Embora reconheçamos a qualidade desse estudo, é importante destacar que sua abordagem apresenta limitações, como a falta de anotações detalhadas dos laudos, restrições no tamanho dos conjuntos de dados, já que emprega apenas 3,069 amostras, e a ausência de validação cruzada em seus experimentos. Essas considerações ressaltam a importância de uma avaliação cuidadosa e abrangente ao comparar os resultados obtidos em diferentes estudos.

Quanto à comparação com os demais trabalhos, nossa pontuação BLEU-1 de 0,612 supera todos os métodos considerados, indicando uma correspondência mais precisa entre os relatórios médicos gerados automaticamente e as referências humanas. Essa superioridade se estende a outras métricas, como BLEU-2, BLEU-3 e BLEU-4, evidenciando a habilidade de nossa abordagem em capturar nuances linguísticas sutis.

Tabela 14 – Resultados do estado-da-arte em geração automática de laudos médicos preliminares para diversas imagens médicas comparados com nossa metodologia.

METHODS	BLEU-1	BLEU-2	BLEU-3	BLEU-4	METEOR	ROUGE-L
(XUE et al., 2024)	0,372*	0,233*	0,154*	0,112*	0,152*	0,286*
(WANG et al., 2018)	0,286*	0,159*	0,104*	0,074*	0,108*	0,226*
(CAO et al., 2023)	**0,799***	**0,692***	**0,634***	**0,589***	-	**0,748***
(HUANG et al., 2019)	0,476*	0,340*	0,238*	0,169*	-	0,347*
(ZHAO et al., 2023)	0,399*	0,158*	0,109*	0,152*	0,275*	
(MOHSAN et al., 2023)	0,532*	0,344*	0,233*	0,158*	0,218*	0,387*
(KOUGIA et al., 2021)	-	-	-	-	-	0,267*
(TSANIYA; FATICHAH; SUCIATI, 2024)	0,363	0,371	0,388	0,412	-	-
(SHAIK; CHERUKURI, 2024)	0,297*	0,230*	0,214*	0,142*	-	0,391
(KONG et al., 2024)	0,280*	0,210*	0,170*	0,140*	0,140*	0,290*
Nosso	0,612	0,552	0,507	0,470	**0,471**	0,633

Em **negrito** os melhores resultados.
Estudos com valores *p* menores que 0.05, quando comparados com nossas métricas recebem *.

Além disso, alcançamos uma pontuação METEOR de 0,471 e uma pontuação ROUGE-L de 0,633, reforçando ainda mais a qualidade de nossos resultados em relação às referências humanas. No entanto, é importante ressaltar que essa comparação tem como objetivo fornecer informações sobre o posicionamento de nosso método em relação a outras pesquisas similares.

Um dos principais diferenciais de nosso trabalho reside em seu foco em resolver um problema anteriormente não explorado relacionado às radiologias de lombo-sacra. Ao utilizar um conjunto de dados especializado e implementar um sistema CAD ponto-a-ponto, oferecemos uma abordagem direcionada especificamente para essa modalidade de imagem médica. Além disso, nossa metodologia incorpora técnicas de pré-processamento e aumento de dados para garantir a robustez e confiabilidade de nosso modelo. Através do ajuste fino profundo de CNNs pré-treinadas e arquiteturas de *Transformer*, nosso método pode aprender características profundas das imagens lombo-sacra, contribuindo para a geração de relatórios médicos mais precisos e informativos. Ao gerar automaticamente relatórios a partir de imagens frontais e laterais de exames lombo-sacra, nossa abordagem adiciona um valor significativo ao processo de diagnóstico, fornecendo uma segunda opinião valiosa para especialistas e potencialmente melhorando os resultados do cuidado ao paciente.

A Tabela 14 apresenta os valores de *p*(KRZYWINSKI; ALTMAN, 2013) obtidos a partir dos testes T de *Student*(KALPIĆ; HLUPIĆ; LOVRIĆ, 2011), comparando nosso método com os métodos estado-da-arte para geração de relatórios médicos a partir de imagens médicas. Os valores de *p* indicam a significância estatística das diferenças entre nosso método e os métodos estado-da-arte nas métricas de avaliação. Valores de *p* menores que 0,05 geralmente indicam que a diferença observada é estatisticamente significativa, ou seja, não é provável que tenha ocorrido por acaso. Com base nos

resultados apresentados, observa-se que nosso método supera ou se equipara a vários métodos estado-da-arte, especialmente nas métricas BLEU-1, BLEU-2, BLEU-3 e METEOR, onde obtivemos os melhores resultados em METEOR e performances competitivas nas demais. Em contrapartida, o método de Cao et al. (2023) apresentou os melhores resultados absolutos em BLEU e ROUGE-L, embora a análise estatística revele que as diferenças entre os dois métodos nem sempre são significativas, indicando que nosso método pode ser uma alternativa válida.

Em termos de tempo computacional, a fase de treinamento raso leva aproximadamente 17 minutos e 56 segundos, enquanto a fase de treinamento profundo dura cerca de 18 minutos e 35 segundos. Carregar o modelo treinado na memória leva cerca de 2 minutos e 26 segundos, e gerar um único relatório médico com o modelo leva aproximadamente 1 segundo. O cálculo das métricas para o conjunto de teste leva cerca de 30 minutos e 6 segundos. O tamanho final do modelo é de 291,1 megabytes. Esses tempos influenciaram significativamente o design dos experimentos, garantindo que os processos de treinamento e inferência fossem eficientes e gerenciáveis dentro de restrições práticas. Além disso, essas limitações de tempo e recursos foram levadas em consideração para garantir uma avaliação rigorosa e detalhada da metodologia proposta.

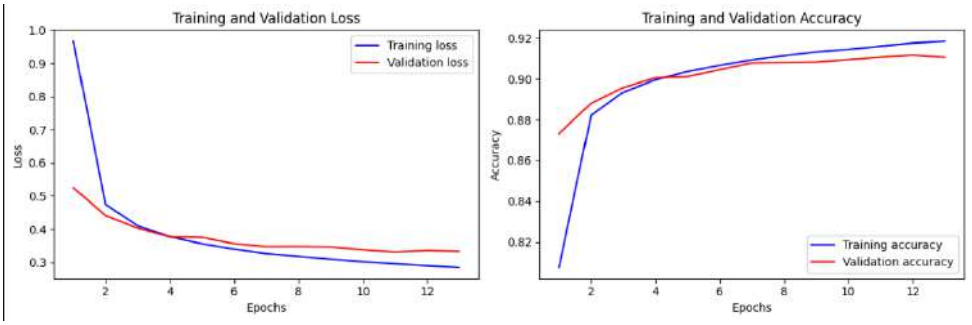
É importante ressaltar que os tempos aqui apresentados refletem o uso da configuração de hardware atual, que consiste em um ambiente de pesquisa com GPUs dedicadas. Em um cenário real, onde várias unidades hospitalares possam acessar simultaneamente o sistema, esses tempos podem mudar drasticamente devido à carga adicional e ao tipo de infraestrutura disponível. Uma eventual implementação em larga escala exigiria adaptações, como a utilização de servidores de empreguem processamento dedicado talvez com arquiteturas distribuídas para garantir que a eficiência e a viabilidade operacional sejam mantidas, mesmo sob uma demanda elevada de acesso.

Essas análises estatísticas e o tempo de execução mostram que, apesar dos desafios, nosso método oferece uma solução eficiente e competitiva para a geração automática de relatórios médicos preliminares, podendo ser implementado em cenários clínicos com grandes volumes de dados, sem comprometer a qualidade dos resultados ou a viabilidade operacional.

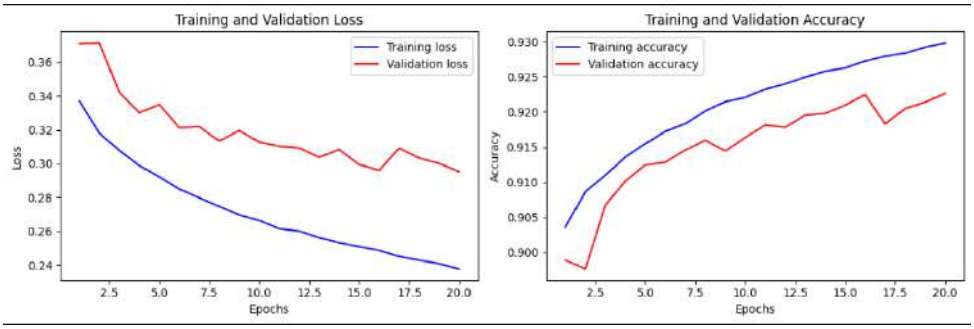
5.2.1 Discussão

Para entendermos os resultados, geramos os gráficos nas Figuras 39a e 39b, que ilustram o desempenho do modelo durante o treinamento e validação. Na Figura 39a, observamos que o modelo apresentou convergência nas métricas de perda e precisão tanto nos conjuntos de treinamento quanto de validação. No entanto, na Figura 39b, notamos uma suavização nas linhas do gráfico no aprendizado do modelo para as

mesmas métricas no conjunto de treinamento em comparação com o conjunto de validação. Essa diferença ocorre porque, na Figura 39a, apenas o *Transformer* está aprendendo as características do conjunto de dados, enquanto na Figura 39b, o treinamento é reiniciado com as CNNs (frontal e lateral) também atualizando seus *kernels* simultaneamente com o *Transformer* já treinado, em um processo de ajuste fino profundo. Apesar dessa adaptação, conseguimos guiar o modelo para capturar as nuances entre as radiografias e os laudos, controlando o sobreajuste.



(a)



(b)

Figura 39 – Gráficos ilustrando o desempenho do modelo durante treinamento e validação. (a) Convergência das métricas de perda e acurácia, com características do conjunto de dados aprendidas apenas pelo *Transformer*. (b) Continuação do aprendizado com ajuste fino profundo, onde tanto as CNNs quanto o *Transformer* aprendem características do conjunto de dados.

Explicar como as CNNs e os *Transformers* alcançam bons resultados é uma tarefa complexa. Métodos de DL podem ser conhecidos como “caixas pretas”; no entanto, a técnica do Grad-CAM permite visualizar regiões importantes após o processo de aprendizagem das CNNs. Combinada com exemplos de laudos gerados pelos *Transformers*, ela pode fornecer informações sobre pontos fortes e áreas que precisam de atenção em trabalhos futuros na nossa metodologia. Para isso, criamos as Figuras 41 e 41, que apresenta exames de lombo-sacra, seus laudos originais, os laudos gerados por nossa metodologia, métricas e os Grad-CAMs para palavras chaves que indicam achados específicos.

Na Figura 40a, apresentamos um exemplo de um exame exibindo tanto o relatório médico quanto o relatório gerado pelo modelo. Observamos que o modelo produziu um

relatório com alto nível de precisão, conforme indicado pelas métricas. Ele capturou nuances do exame, como leve desvio para a direita e rarefação óssea. No entanto, deixou passar duas vértebras mencionadas no relatório original, T12 e L1. Além disso, identificou uma redução, mas apenas na região L5-S1. Quanto à esclerose, o modelo detectou um espectro maior de vértebras em comparação com o relatório original. Um comportamento semelhante também é observado na Figura 40b.

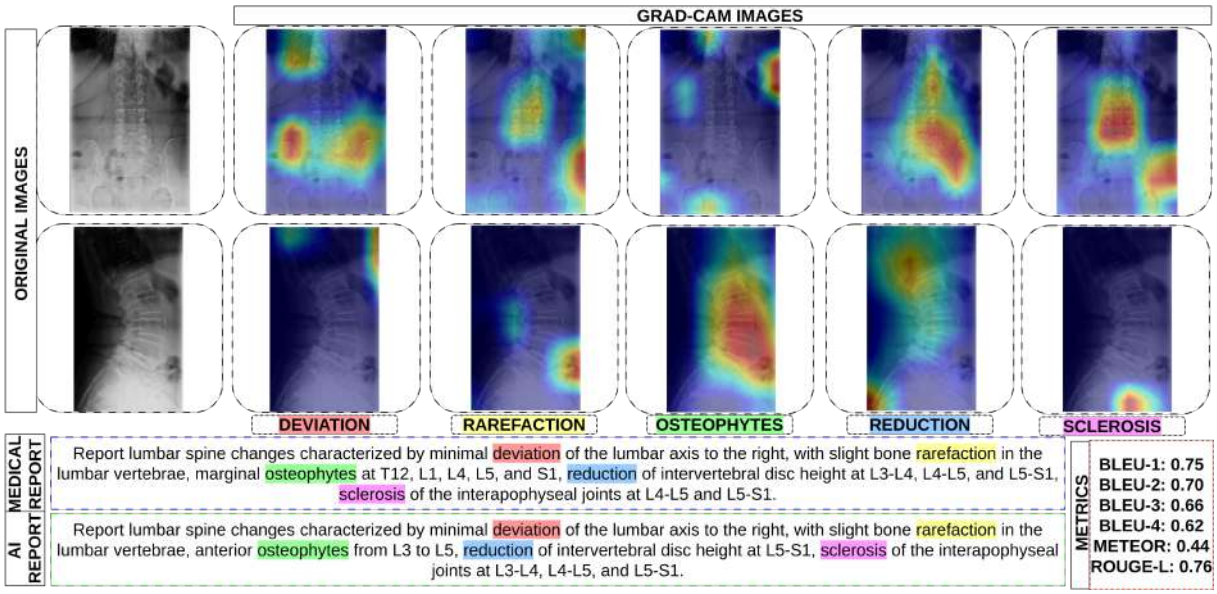
A Figura 41a oferece detalhes distintos em comparação com as Figuras 40a e 40b. Enquanto o modelo nessas duas figuras gerou relatórios muito semelhantes aos originais, diferindo apenas no grau dos problemas identificados, na Figura 41a, o modelo apresenta um relatório mais complexo e informativo que o original. Por exemplo, destaca-se a tentativa do modelo de evidenciar a presença de problemas de osteófitos da L3 à L5 e também de esclerose interapofisária em vértebras específicas, enquanto o relatório médico não menciona esses problemas. Da mesma forma, a Figura 41b exibe um comportamento semelhante ao da Figura 41a; no entanto, aqui o modelo identifica ainda mais achados específicos que o laudo original.

Outro aspecto relevante é que, frequentemente, as métricas podem oferecer resultados que não correspondem ao relatório gerado, já que o modelo pode expressar algo semelhante ao relatório real, mas com um conjunto diferente de palavras e expressões. Isso pode resultar em uma avaliação inferior, destacando o quão desafiadora é a tarefa de quantificar textos.

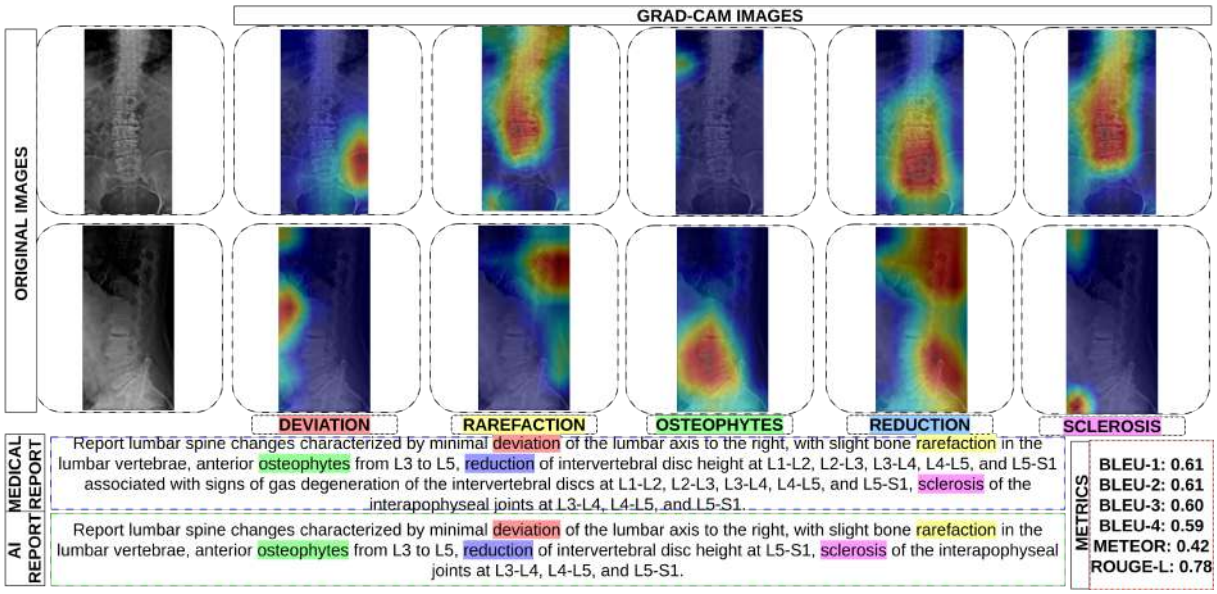
5.3 Resultados Para Geração Automática de Laudos Preliminares em Pododáctilos

Nesta seção, apresentamos os resultados obtidos pela metodologia de modelos generativos para laudos médicos em radiografias pododáctilos, conforme mostrado na Tabela 15, seguido por uma breve discussão. Em seguida, comparamos nossos resultados com o estado-da-arte na Tabela 16 para entender melhor como nosso método se posiciona em relação a outras pesquisas semelhantes.

A metodologia proposta demonstrou um desempenho promissor na geração automática de laudos preliminares de raios-X de pododáctilos. Os resultados das métricas BLEU, METEOR e ROUGE-L para os rótulos ANORMAL, NORMAL e *TOGETHER* indicam que o modelo é particularmente eficaz em casos normais, com pontuações significativamente mais altas (BLEU-1 de 0,653 e METEOR de 0,548) em comparação aos casos anormais (BLEU-1 de 0,426 e METEOR de 0,326). Esta diferença de desempenho pode ser atribuída à complexidade inerente em descrever achados anormais e à escassez de amostras para determinadas patologias, o que dificulta o processo de aprendizado. A baixa variabilidade observada nas métricas, indicada pelos desvios padrão, sugere que o

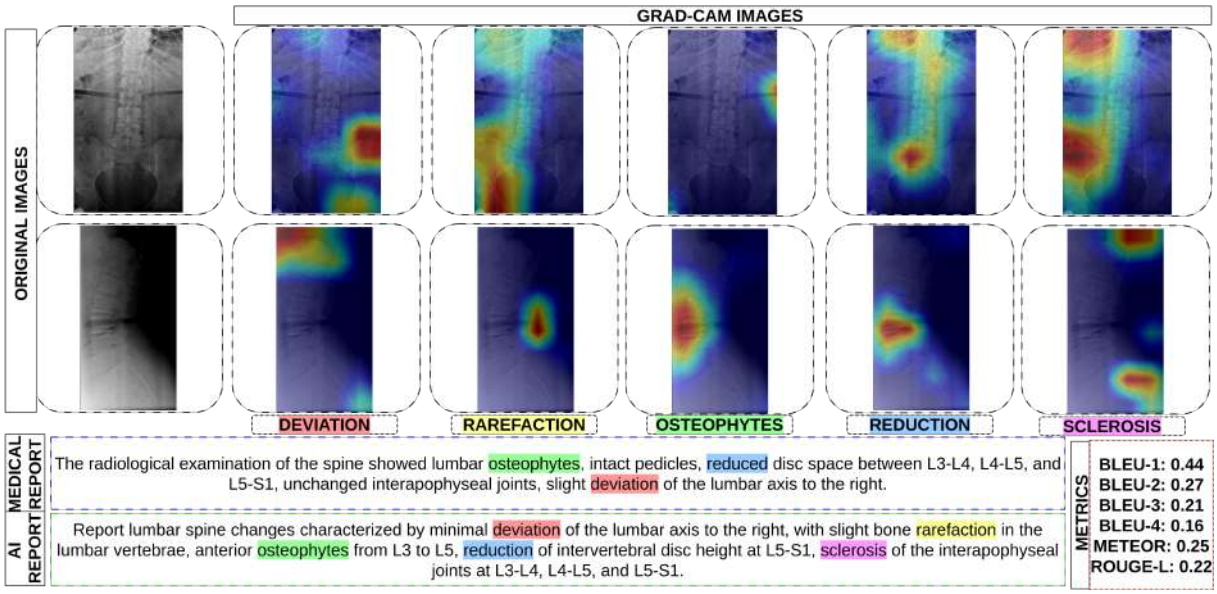


(a)

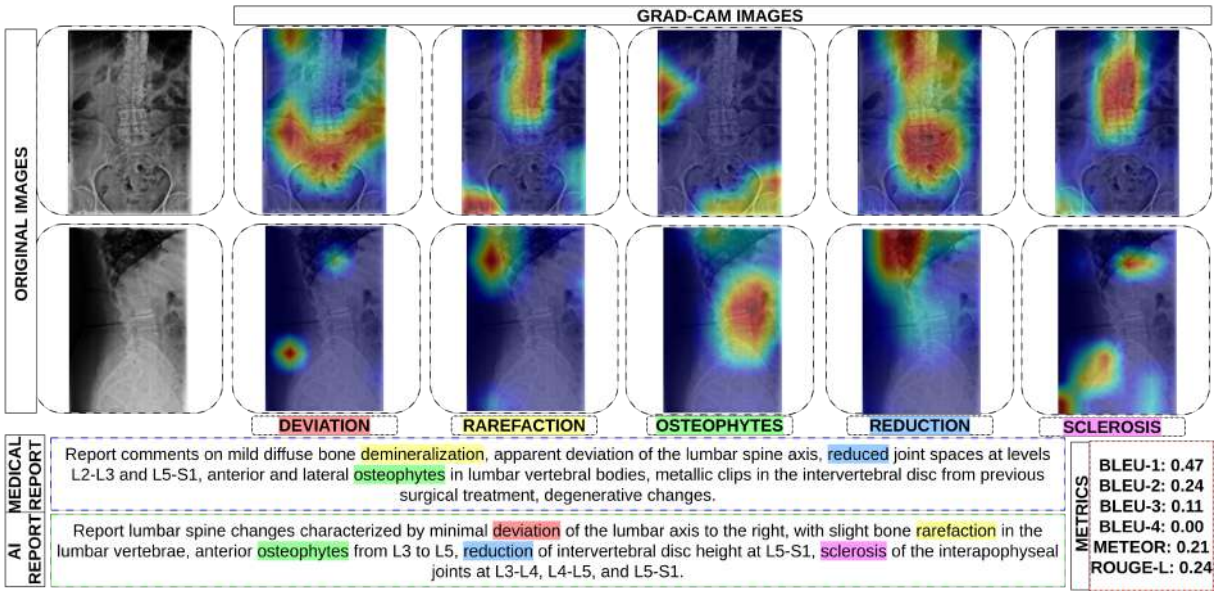


(b)

Figura 40 – Exemplos de exames (a, b), relatórios médicos gerados pelo nosso método e respectivas métricas de avaliação. As cores indicam no texto o Grad-CAM específico para determinadas palavras-chave sendo vermelho para "desvio", amarelo para "refração", verde para "osteófitos", azul para "redução"e rosa para "esclerose".



(a)



(b)

Figura 41 – Exemplos de exames (a, b), relatórios médicos gerados pelo nosso método e respectivas métricas de avaliação. As cores indicam no texto o Grad-CAM específico para determinadas palavras-chave sendo vermelho para "desvio", amarelo para "refração", verde para "osteófitos", azul para "redução"e rosa para "esclerose".

modelo apresenta consistência ao longo dos diferentes *folds* do conjunto de validação. No entanto, a maior variabilidade nas pontuações BLEU-2 a BLEU-4 para os casos normais indica espaço para melhorias no manejo de sequências mais longas. No geral, a capacidade do modelo de gerar laudos precisos tanto para casos normais quanto anormais reforça sua aplicabilidade no mundo real. Para avanços futuros, recomenda-se a inclusão de mais dados de treinamento e a aplicação de técnicas avançadas de ajuste fino para aprimorar a precisão e a confiabilidade do modelo, contribuindo para diagnósticos médicos mais eficientes e precisos.

Tabela 15 – Resultados médios das métricas para cada *fold*, juntamente com o desvio padrão.

Classess	**BLEU-1**	**BLEU-2**	**BLEU-3**	**BLEU-4**	**METEOR**	**ROUGE-L**
AN	0,426±0,030	0,340±0,031	0,303±0,034	0,282±0,037	0,326±0,022	0,271±0,027
N	0,653±0,072	0,570±0,102	0,509±0,094	0,503±0,094	0,548±0,068	0,504±0,078
T	0,516±0,018	0,432±0,026	0,386±0,020	0,370±0,019	0,414±0,018	0,364±0,021

AN é Anormal, N é Normal e T é Todos

Os resultados observados na Tabela 15 sugerem que a metodologia obteve resultados promissores, o que pode ser atribuído à abordagem adotada, que inclui etapas de pré-processamento tanto para imagens quanto para texto, juntamente com a integração de quatro CNNs pré-treinadas com transformadores para interpretação de radiografias e geração automática de laudos preliminares. Essa metodologia permite que o modelo capture nuances textuais presentes em imagens de radiografias de pododáctilos. Além disso, a tabela indica que o modelo apresenta melhor desempenho para relatórios sem anormalidades em comparação com aqueles considerados anormais. Essa informação é fundamental para orientar decisões sobre como a metodologia pode auxiliar os profissionais médicos.

Uma análise mais aprofundada pode ser feita a partir da Figura 42, onde o gráfico de caixas revela uma distribuição consistente das métricas de avaliação (BLEU-1, BLEU-2, BLEU-3, BLEU-4, METEOR e ROUGE-L) ao longo das 10 dobras. A mediana relativamente boa em todas as métricas sugere consistência na qualidade das previsões, enquanto a presença de *outliers* indica variações em determinadas iterações da validação cruzada. No entanto, a distribuição geral das métricas, incluindo METEOR e ROUGE-L, permanece estável entre as dobras, sugerindo uma consistência geral na capacidade do modelo de gerar relatórios médicos automáticos. Esses resultados indicam que nosso método mantém um desempenho uniforme e confiável, estabelecendo uma base sólida para a confiabilidade e robustez do modelo proposto.

Ao examinarmos o *box plot*, observamos que tanto as classes “Anormal” quanto “Normal” exibem distribuições que podem ser consideradas semelhantes em métricas ao longo das dobras. No entanto, é importante destacar novamente que os resultados para exames da classe anormal apresentam um desempenho ligeiramente inferior em

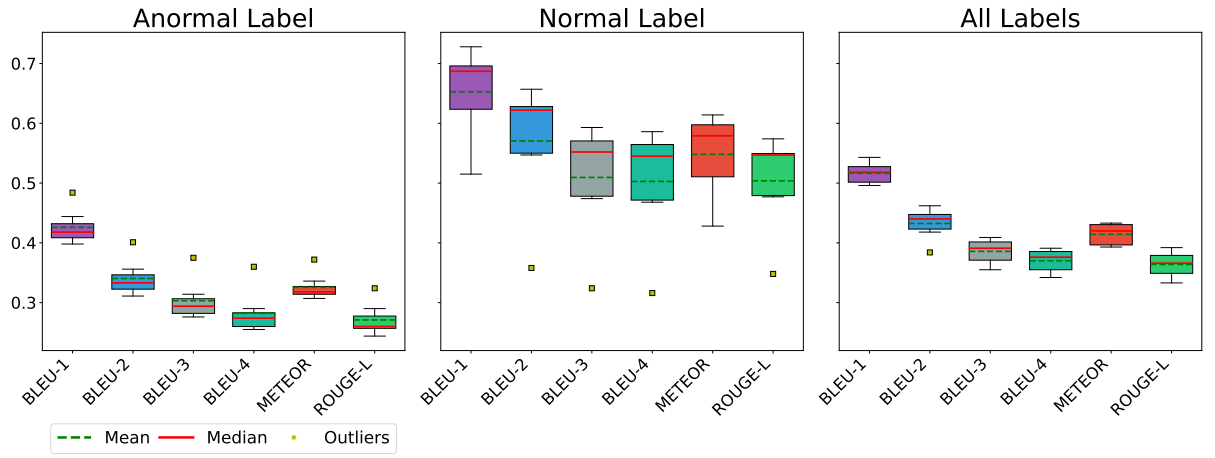


Figura 42 – *Box-plot* dos resultados da geração automática de laudos médicos preliminares com as métricas BLEU-1, BLEU-2, BLEU-3, BLEU-4, METEOR e ROUGE-L, considerando os 10 *folds*, para as classes “Anormal” e “Normal”, bem como para a combinação de ambas (”Todas as Classes”).

comparação com a classe normal, com diferenças na dispersão dos dados. A classe “Anormal” apresenta medianas ligeiramente mais baixas e intervalos interquartis mais amplos em comparação com a classe “Normal”. Isso reforça que, embora o desempenho geral permaneça consistente, pode haver ligeiramente mais variabilidade nas previsões para radiografias anormais.

Ao analisarmos a combinação de ambas as classes, observamos que os resultados médios situam-se entre os das classes anormal e normal. O modelo demonstra robustez geral, sugerindo que ele é promissor para a geração de laudos médicos automáticos em diversos cenários clínicos. No entanto, devido à maior variabilidade observada nos achados específicos da classe anormal, o modelo mostrou-se menos eficiente para este cenário. Esta informação é crucial para a tomada de decisão sobre a utilização do modelo em um ambiente real. Adicionalmente, indica que a inclusão de um número maior de amostras da classe anormal poderia contribuir ainda mais na robustez dos resultados, tornando o modelo mais confiável e eficaz para uma gama mais ampla de diagnósticos.

A seguir, realizamos uma análise comparativa entre nosso método proposto e os métodos do estado-da-arte, apresentada na Tabela 16. É importante ressaltar que essa comparação é feita apenas com o propósito de entendermos em que etapa se encontra nosso método na geração de laudos automáticos para exames em sistemas CAD. Deve-se entender que uma comparação direta não pode ser feita, pois neste estudo resolvemos um problema diferente e inédito apontado em nosso estado-da-arte.

Tabela 16 – Resultados estado-da-arte em geração automática de laudos médicos preliminares para diversos tipos de imagens médicas comparados com nossa metodologia.

METODOS	BLEU-1	BLEU-2	BLEU-3	BLEU-4	METEOR	ROUGE-L
(XUE et al., 2024)	*0,372	*0,233	*0,154	*0,112	*0,152	0,286
(WANG et al., 2018)	*0,286	*0,159	*0,104	*0,074	*0,108	*0,226
(CAO et al., 2023)	***0,799**	***0,692**	***0,634**	***0,589**	-	***0,748**
(HUANG et al., 2019)	0,476	0,340	*0,238	*0,169	-	0,347
(ZHAO et al., 2023)	*0,399	*0,158	*0,109	*0,152	*0,275	-
(MOHSAN et al., 2023)	0,532	0,344	*0,233	*0,158	*0,218	0,387
(KOUGIA et al., 2021)	-	-	-	-	-	0,267
(TSANIYA; FATICHAH; SUCIATI, 2024)	*0,363	0,371	0,388	0,412	-	-
(SHAIK; CHERUKURI, 2024)	*0,297	*0,230	*0,214	*0,142	-	0,391
(KONG et al., 2024)	*0,280	*0,210	*0,170	*0,140	*0,140	0,290
Nosso	0,516	0,432	0,386	0,370	**0,414**	0,364

negrito os melhores resultados. Estudos com valores *p* menores que 0,05, quando comparados com nossas métricas recebem *.

Nosso método se destaca consistentemente em várias métricas de avaliação, demonstrando seu potencial na geração de laudos médicos para exames radiográficos de Polidactilia. No entanto, ao ser comparado ao estudo de CAO et al., que investigou o uso de imagens de endoscopia, observamos resultados superiores em métricas específicas. É importante ressaltar que, ao contrário de nossa abordagem, o estudo de Cao et al. (2023). não fornece uma avaliação abrangente de todas as métricas usadas em nosso estudo. Embora reconheçamos a qualidade deste estudo, é crucial mencionar suas limitações, como a ausência de anotações detalhadas nos laudos, restrições no tamanho do conjunto de dados, com apenas 3,069 amostras, e a falta de validação cruzada em seus experimentos. Essas considerações destacam a importância de uma avaliação completa e abrangente ao comparar os resultados de diferentes estudos.

Dado que o método proposto não alcançou os melhores resultados na literatura, é necessário entender que o objetivo desta comparação não é determinar a melhor metodologia, pois tal comparação seria injusta, já que os métodos abordam problemas diferentes. No entanto, essa comparação é importante para entender o nível de desempenho do nosso método em comparação com outros que também resolvem o problema de geração automática de relatórios médicos preliminares. Dito isso, as vantagens do nosso método em comparação com Cao et al. (2023) são 1) um conjunto de dados maior e mais heterogêneo; 2) uma avaliação dos resultados com mais métricas, o que permite uma melhor análise e interpretação dos nossos resultados. Nosso estudo contribui para o campo emergente da IA na medicina, explorando novas metodologias para melhorar a interpretação automática de imagens médicas.

Em relação à comparação com outros trabalhos, nossa pontuação BLEU-1 de 0,516 supera a apresentada por (MOHSAN et al., 2023) e é superior a todos os demais métodos considerados na Tabela 16, indicando uma correspondência precisa entre os relatórios médicos gerados automaticamente e as referências humanas. Além disso,

obtivemos resultados comparáveis em outras métricas, como BLEU-2, BLEU-3 e BLEU-4, evidenciando a capacidade de nossa abordagem em capturar nuances linguísticas sutis. Alcançamos uma pontuação METEOR de 0,417, superior a todas as outras metodologias, e uma pontuação ROUGE-L de 0,364, reforçando ainda mais a qualidade de nossos resultados em relação às referências humanas. Esta análise demonstra que, ao resolver o problema da geração automática de laudos médicos preliminares para exames de pododáctilos, conseguimos resultados comparáveis e, em alguns casos, superiores em determinadas métricas em comparação com o estado-da-arte, evidenciando a eficácia de nossa metodologia.

A Tabela 16 apresenta os valores de *p* (KRZYWINSKI; ALTMAN, 2013) obtidos a partir dos testes *T de Student* (KALPIĆ; HLUPIĆ; LOVRIĆ, 2011), comparando nosso método com os métodos do estado-da-arte para geração de relatórios médicos a partir de imagens médicas. A tabela 16 mostra os resultados das métricas para cada um dos métodos do estado-da-arte comparados com nosso método. Os valores de p indicam a significância estatística das diferenças entre nosso método e os métodos estado-da-arte nas métricas de avaliação. Valores de *p* menores que 0,05 geralmente indicam que a diferença observada é estatisticamente significativa, ou seja, não é provável que tenha ocorrido por acaso. Com base nos resultados apresentados, observa-se que nossos métodos superam ou se equiparam a vários métodos estado-da-arte, especialmente nas métricas BLEU-1, BLEU-2, BLEU-3, e METEOR, onde obtivemos os melhores resultados em METEOR e performances competitivas nas demais. Isso sugere que nosso modelo não apenas é robusto, mas também oferece uma alternativa viável e eficaz na geração automática de laudos médicos preliminares. Em contrapartida, o método deCao et al. (2023) apresentou os melhores resultados absolutos em BLEU e ROUGE-L, embora a análise estatística revele que as diferenças entre os dois métodos não são sempre significativas, indicando que nosso método pode ser uma alternativa válida. Essas análises estatísticas são essenciais para confirmar a superioridade ou equivalência dos métodos, garantindo que as melhorias observadas sejam consistentes e não atribuíveis ao acaso.

5.3.1 Discussão

Nesta seção, apresentamos os resultados do nosso método para geração automática de laudos médicos preliminares de pododáctilos, utilizando CNNs para extrair características das imagens e *Transformers* para interpretar e gerar os relatórios. Começamos ilustrando o desempenho de treinamento e validação do nosso modelo por meio de diversos gráficos. Além disso, fornecemos exemplos de relatórios gerados juntamente com as respectivas imagens médicas, incluindo visualizações Grad-CAM para melhorar a interpretabilidade dos resultados (VIEIRA et al., 2021; SIRIPATTANADILOK;

SIRIBORVORNRATANAKUL, 2024).

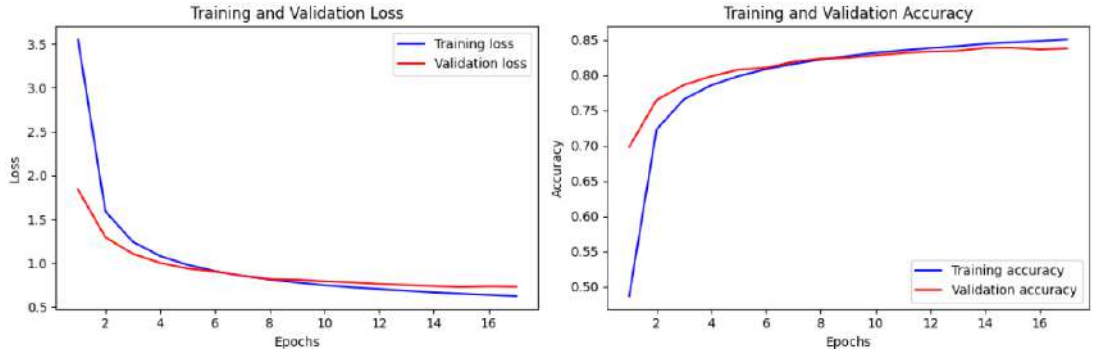
Para esta seção, selecionamos o modelo que apresentou resultados médios entre os 10 *folds*. Optamos por este enfoque para mitigar vieses amostrais e garantir uma representação robusta da generalização do modelo. Escolher um modelo com desempenho médio é crucial pois ele representa uma média do desempenho em diferentes conjuntos de validação, proporcionando uma avaliação mais realista da capacidade do método em lidar com novos dados. Isso não só aumenta a confiança na replicabilidade dos resultados, mas também oferece uma visão mais equilibrada do potencial do modelo em cenários práticos de aplicação clínica.

Para um entendimento mais profundo dos resultados, apresentamos os gráficos nas Figuras 43a e 43b, que ilustram o desempenho do modelo durante os períodos de treinamento e validação. A Figura 43a mostra que o modelo alcançou uma convergência satisfatória em relação às métricas de perda e acurácia nos conjuntos de treinamento e validação, com a perda consistentemente abaixo de 1,0 e a acurácia acima de 0,80. Nesta etapa, apenas o *Transformer* aprendeu características do conjunto de dados. A Figura 43 permite observar que, na segunda etapa de treinamento, onde as camadas convolucionais das quatro CNNs também foram permitidas a aprender junto com a continuação do treinamento do *Transformer*, o modelo não conseguiu mitigar o sobreajuste, resultando em uma interrupção do treinamento. Apesar desse comportamento, na segunda etapa, conseguimos orientar o modelo a aprender as nuances entre as radiografias e os relatórios, controlando o sobreajuste, como observado na Seção 5.3.

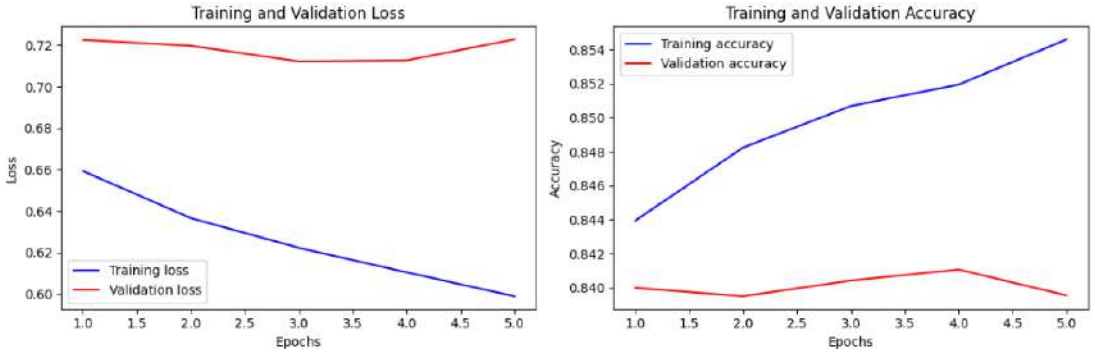
Para elucidar como a combinação das CNNs com os *Transformers* alcançou os resultados apresentados neste estudo, empregamos a técnica Grad-CAM (*Gradient-weighted Class Activation Mapping*) (SELVARAJU et al., 2016) para visualizar as regiões das imagens consideradas importantes pelas CNNs durante o processo. Em seguida, combinamos os laudos gerados pelos *Transformers* com os Grad-CAMs, proporcionando informações adicionais sobre nossa metodologia. Desta forma, elaboramos a Figura 44, que exibe um exame de pododáctilos, seus laudos originais, os laudos produzidos por nossa metodologia, as métricas calculadas e os Grad-CAMs para palavras-chave geradas pela metodologia.

Ao analisarmos a Figura 44a, percebemos que o modelo apresentou um desempenho promissor, com pontuações BLEU-4 de 0,63, METEOR de 0,45 e ROUGE-L de 0,73, mostrando uma forte correspondência com laudos médicos reais. Embora o modelo não tenha detectado um desvio em valgo do retropé/médiopé em estágio inicial, ele foi preciso na descrição de características normais. A análise dos Grad-CAMs gerados pelas quatro CNNs para cada imagem do exame revelou que o modelo focou em áreas dos pés usadas por especialistas para gerar laudos.

A Figura 44b mostra que o modelo obteve pontuações de BLEU-4 de 0,60,



(a)



(b)

Figura 43 – Gráficos ilustrando o desempenho do modelo durante treinamento e validação. (a) Convergência das métricas de perda e acurácia, com características do conjunto de dados aprendidas apenas pelo *Transformer*. (b) Continuação do aprendizado com ajuste fino profundo, onde tanto as CNNs quanto o *Transformer* aprendem características do conjunto de dados.

METEOR de 0,53 e ROUGE-L de 0,91, indicando uma forte correspondência com laudos médicos reais. Ambos os relatórios mencionam textura óssea normal, superfícies e espaços articulares íntegros, e redução do cavo plantar, demonstrando que o modelo capturou com precisão as características clínicas normais. Embora as pontuações sugiram alta concordância, há variação nas palavras devido à origem diversa do conjunto de dados. A análise dos Grad-CAMs revela que o modelo foca em áreas específicas dos pés, indicando que procura por achados específicos.

Na última Figura 44c de Grad-CAMs, podemos ver que o modelo obteve pontuações baixas em BLEU-4 de 0,00, METEOR de 0,10 e ROUGE-L de 0,18, indicando uma baixa correspondência com os laudos médicos reais. O médico fez uma análise breve, não detectando nenhum achado específico, enquanto o modelo realizou uma análise mais detalhada, identificando a presença de hálux valgo e um pino cirúrgico que o médico não mencionou. Este exemplo ilustra a dificuldade de quantificar resultados de *image-to-text*, pois mesmo quando o modelo gera um laudo preciso, o uso de palavras e expressões diferentes pode resultar em métricas baixas.

A Figura 45a apresenta os resultados da análise do modelo sobre imagens

de raios-X dos pés direito e esquerdo, com a visualização das regiões importantes identificadas pela técnica Grad-CAM. Os relatórios gerados pelo modelo, quando comparados aos relatórios médicos, mostraram uma correspondência significativa, conforme evidenciado pelas métricas BLEU-4 de 0.76, METEOR de 0.50 e ROUGE-L de 0.81. Notamos que o modelo foi preciso na identificação de características normais dos ossos e articulações, assim como na detecção de hallux valgus em ambos os pés. Contudo, houve uma discrepância na identificação de sinais de osteoartrite no pé esquerdo, não mencionados pelo modelo. Isso sugere que, embora o modelo tenha um bom desempenho geral, ainda há espaço para aprimoramentos na detecção de anomalias mais sutis, como a osteoartrite.

Na Figura 45b, os resultados mostram que o modelo conseguiu capturar várias características normais do pé direito, com métricas BLEU-4 de 0,59, METEOR de 0,41 e ROUGE-L de 0,72. Contudo, houve divergências consideráveis no pé esquerdo, onde o modelo não identificou o hallux valgus, sinais de osteoartrite e a presença de um parafuso metálico que fixa o metatarso e a falange proximal do primeiro dedo. A análise dos Grad-CAMs revela que as atenções do modelo estão concentradas nas partes importantes das imagens, indicando que o modelo reconhece regiões relevantes, mas ainda falha na identificação precisa de certas anomalias no pé esquerdo. Isso sugere que, embora o modelo esteja focando corretamente nas áreas importantes, pode haver a necessidade de ajustes adicionais para melhorar a sensibilidade e a precisão na detecção de características anômalas.

A Figura 45c apresenta um desempenho relativamente baixo do modelo, com métricas BLEU-4 de 0,08, METEOR de 0,17 e ROUGE-L de 0,31. O relatório gerado pelo modelo falhou em identificar sinais de osteoartrite talonavicular e hallux valgus presentes no relatório médico. Uma observação relevante é o aspecto embaçado das imagens deste exame, com uma coloração esbranquiçada fora dos pés, o que pode ter contribuído para os resultados insatisfatórios do modelo. A análise dos Grad-CAMs revela que a imagem possui características não comuns ao restante do conjunto de dados, dificultando a concentração do modelo em características importantes e afetando negativamente a performance.

Os escores BLEU, METEOR e ROUGE-L alcançados no presente estudo são indicadores quantitativos da correspondência entre os relatórios gerados pelo modelo e os relatórios médicos reais. Na prática clínica, esses escores são de grande relevância, pois fornecem uma medida objetiva da precisão e qualidade das descrições automatizadas em relação às observações feitas por radiologistas experientes. Um escore BLEU-4 de 0,76, por exemplo, sugere que o modelo é altamente eficaz em replicar a linguagem utilizada em relatórios médicos, o que pode acelerar a elaboração de laudos e reduzir a carga de trabalho dos radiologistas. Da mesma forma, os escores METEOR e ROUGE-L, que

atingiram valores de até 0,50 e 0,81 respectivamente, indicam uma forte concordância na estrutura e conteúdo dos relatórios, ressaltando a capacidade do modelo em capturar detalhes críticos das imagens radiológicas. No entanto, a variabilidade dos escores em diferentes cenários clínicos evidencia a necessidade de contínuos aprimoramentos, especialmente na detecção de anomalias sutis.

É importante notar, contudo, que essas métricas nem sempre representam uma análise realista da capacidade do modelo. A variação na escolha de palavras e expressões pode resultar em escores relativamente baixos, mesmo quando o modelo gera laudos tão precisos quanto os dos médicos. Por exemplo, o uso de sinônimos ou de uma estrutura de frase diferente pode diminuir os escores, apesar de o conteúdo clínico ser equivalente. Isso indica que, enquanto as métricas BLEU, METEOR e ROUGE-L são úteis para uma avaliação inicial, uma análise mais profunda e qualitativa é necessária para avaliar plenamente a eficácia do modelo. Portanto, a integração plena dessa tecnologia na prática clínica deve considerar tanto a análise quantitativa quanto a qualitativa para garantir a precisão e a utilidade dos laudos gerados.

Em termos de tempo computacional, a fase de treinamento raso leva aproximadamente 33 minutos e 51 segundos, enquanto a fase de treinamento profundo dura cerca de 35 minutos e 9 segundos. Carregar o modelo treinado na memória leva cerca de 4 minutos e 52 segundos, e gerar um único relatório médico com o modelo leva aproximadamente 1 segundo. O cálculo das métricas para o conjunto de teste leva cerca de 40 minutos e 12 segundos. O tamanho final do modelo é de 582,2 megabytes. Esses tempos influenciaram significativamente o design dos experimentos, garantindo que os processos de treinamento e inferência fossem eficientes e gerenciáveis dentro de restrições práticas. Além disso, essas limitações de tempo e recursos foram levadas em consideração para garantir uma avaliação rigorosa e detalhada da metodologia proposta.

Novamente, é importante ressaltar que os tempos aqui apresentados refletem o uso da configuração de hardware atual, que consiste em um ambiente de pesquisa com GPUs dedicadas. Em um cenário real, onde várias unidades hospitalares possam acessar simultaneamente o sistema, esses tempos podem mudar drasticamente devido à carga adicional e ao tipo de infraestrutura disponível. Uma eventual implementação em larga escala exigiria adaptações, como a utilização de servidores de empreguem processamento dedicado talvez com arquiteturas distribuídas para garantir que a eficiência e a viabilidade operacional sejam mantidas, mesmo sob uma demanda elevada de acesso.

Essa discussão sugere que, embora o modelo apresente um desempenho promissor na descrição de características clínicas normais e na identificação de achados específicos, há necessidade de ajustes adicionais para melhorar a detecção de anomalias iniciais e mitigar o impacto da variação linguística. Com esses aprimoramentos, como a

adição de novas amostras de exames ao conjunto de dados, o modelo tem grande potencial para ser uma ferramenta valiosa na prática clínica, auxiliando na interpretação de raios-X de pododáctilos com precisão e eficiência.

5.4 Conclusões

Neste estudo, desenvolvemos e avaliamos uma metodologia baseada em modelos de inteligência artificial generativa para a criação automática de laudos médicos a partir de radiografias da coluna lombo-sacra e dos pododáctilos. Nosso objetivo principal foi alcançado ao criar um método preciso e especializado para gerar laudos preliminares, auxiliando médicos no processo diagnóstico e promovendo o avanço dos sistemas CAD aplicados à radiologia.

A interpretabilidade dos resultados foi aprimorada com a aplicação da técnica Grad-CAM, que permitiu visualizar as regiões das imagens que mais contribuíram para a geração dos laudos. Isso fornece maior transparência ao modelo e aumenta a confiança dos profissionais de saúde ao utilizarem a ferramenta como suporte diagnóstico.

Observamos que os resultados para as radiografias de lombo-sacra foram superiores aos obtidos para os pododáctilos. Essa diferença pode ser atribuída à menor quantidade de dados disponíveis para as radiografias de pododáctilos, o que limita o modelo em termos de aprendizado de padrões complexos e afeta a precisão dos laudos gerados. Essa limitação ressalta a importância de conjuntos de dados amplos e diversificados para o treinamento de modelos de inteligência artificial em contextos médicos.

Em relação aos objetivos propostos na introdução, podemos afirmar que o segundo objetivo específico foi plenamente atingido. Desenvolvemos e treinamos um modelo de IA generativa para auxiliar na geração de laudos médicos descritivos e diagnósticos a partir de radiografias da coluna lombo-sacra e dos pododáctilos. Esses laudos servem como suporte à análise dos especialistas, oferecendo uma segunda opinião que contribui para a precisão e eficiência dos diagnósticos médicos.

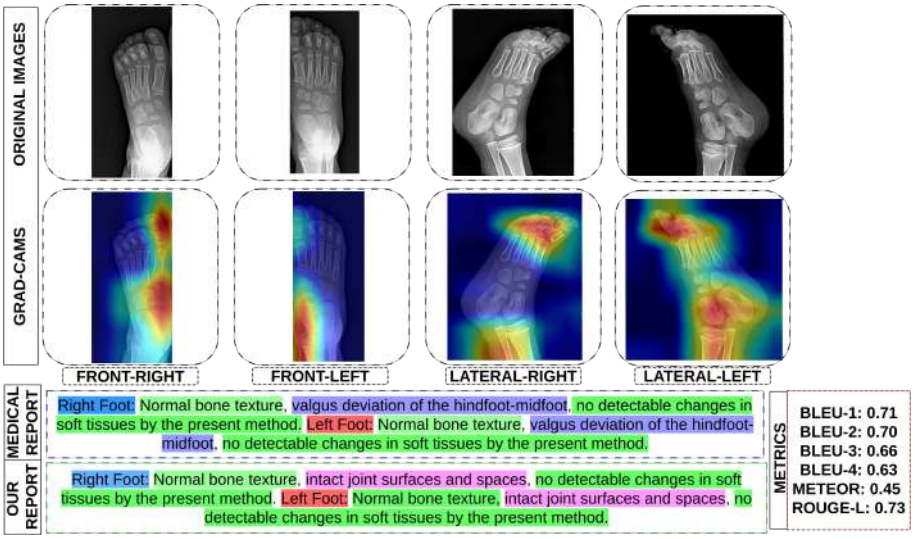
Além disso, o terceiro objetivo específico, que consistia em avaliar o desempenho dos modelos desenvolvidos, comparando-os com métodos existentes e utilizando métricas específicas para validar a precisão, confiabilidade e eficiência na qualidade dos laudos, também foi cumprido. As comparações realizadas com o estado da arte mostraram que nosso modelo é competitivo e apresenta resultados promissores, reforçando sua relevância no contexto clínico.

No entanto, reconhecemos algumas limitações em nosso trabalho. A geração de laudos para exames com anomalias foi menos precisa do que para os casos normais, evidenciando a necessidade de aumentar o número de amostras de exames anômalos

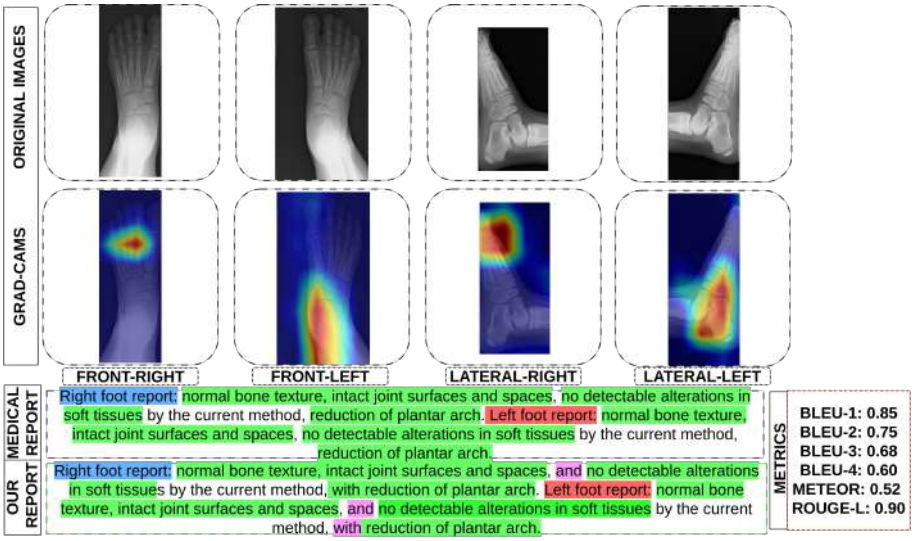
para melhorar o aprendizado do modelo. A aplicação da metodologia ficou restrita às radiografias da coluna lombo-sacra e dos pododáctilos, limitando sua generalização para outras modalidades de imagens médicas.

Para superar essas limitações, sugerimos, como trabalhos futuros, a incorporação de conjuntos de dados mais amplos e diversificados, especialmente com maior representatividade de exames anômalos. Além disso, explorar arquiteturas de CNN alternativas para a extração de características e investigar diferentes configurações de hiperparâmetros em Transformers pode contribuir para aprimorar o desempenho do modelo. Expandir o escopo do modelo para incluir diferentes modalidades de exames clínicos ampliaria sua utilidade em um cenário clínico mais diversificado.

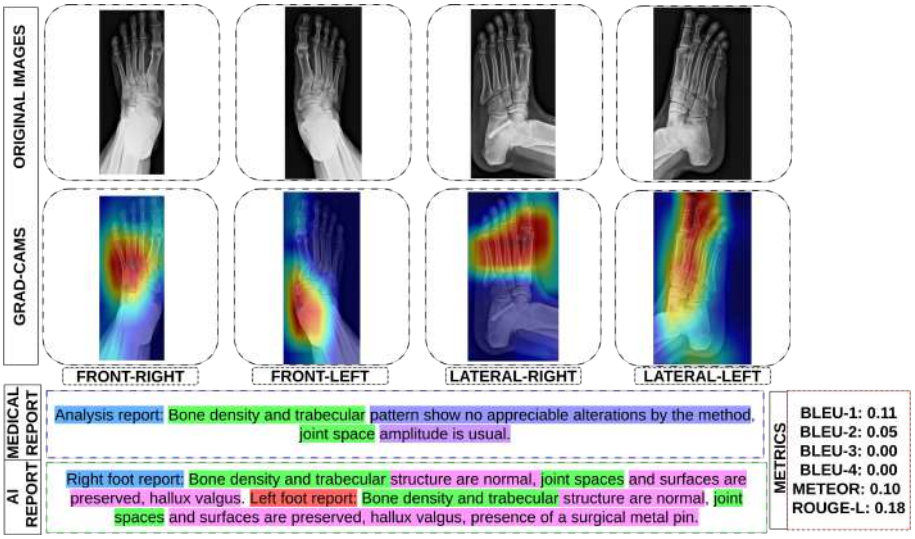
Em suma, este estudo oferece uma contribuição significativa para o campo emergente da inteligência artificial na medicina, demonstrando o potencial de abordagens avançadas de aprendizado de máquina para otimizar o processo de interpretação de imagens médicas e apoiar decisões clínicas em diferentes contextos radiológicos. Acreditamos que a integração de modelos de IA generativa em sistemas CAD pode aprimorar a precisão e eficiência diagnóstica, reduzir a carga de trabalho dos profissionais de saúde e fornecer suporte valioso na tomada de decisão clínica, promovendo um ambiente clínico mais seguro e eficaz.



(a)

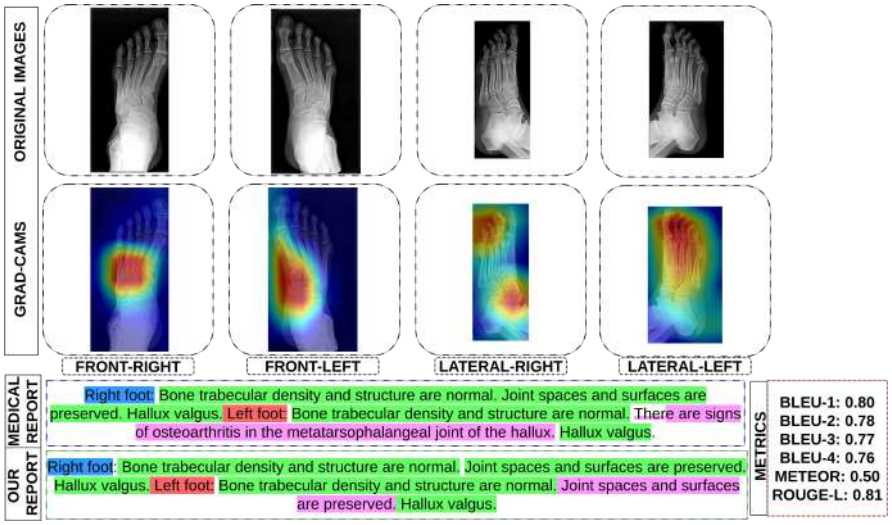


(b)

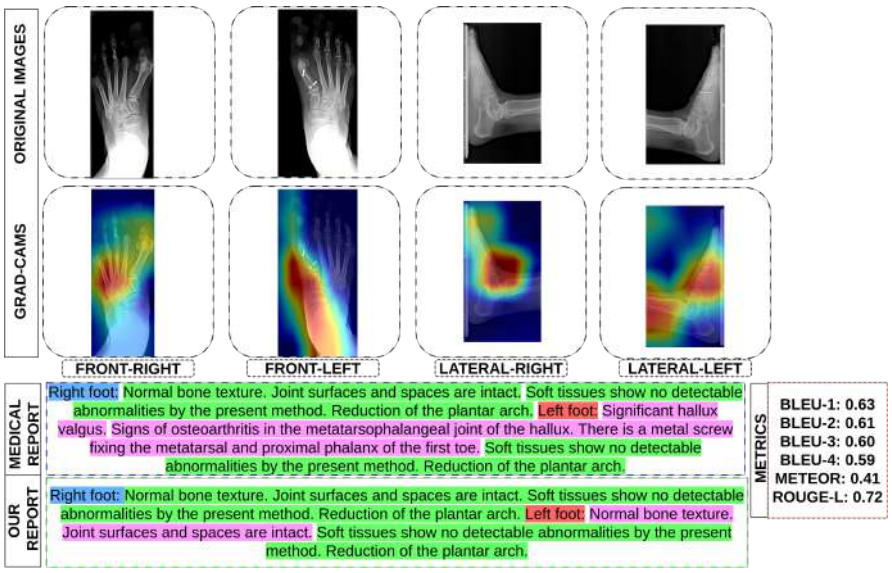


(c)

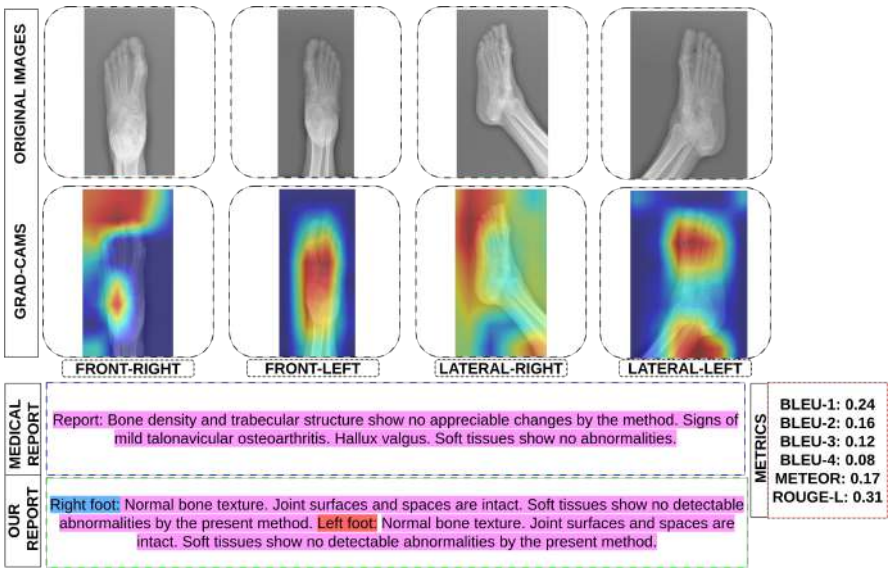
Figura 44 – Exemplos de exames (a - c), relatórios médicos gerados pelo nosso método e métricas de avaliação. Referências do pé direito em azul, do pé esquerdo em vermelho. Texto verde indica correspondências exatas, texto roxo indica omissões, e texto rosa destaca adições ou discrepâncias do modelo.



(a)



(b)



(c)

Figura 45 – Exemplos de exames (a - c), relatórios médicos gerados pelo nosso método e métricas de avaliação. Referências do pé direito em azul, do pé esquerdo em vermelho. Texto verde indica correspondências exatas, texto roxo indica omissões, e texto rosa destaca adições ou discrepâncias do modelo.

6 Considerações Finais

A radiografia permanece como uma ferramenta indispensável na medicina moderna, amplamente utilizada para diagnosticar uma variedade de condições médicas, graças à sua capacidade de fornecer imagens detalhadas baseadas na absorção diferencial dos raios-X por tecidos de diferentes densidades. No entanto, a análise dessas imagens apresenta desafios significativos, incluindo a identificação de anomalias sutis e a interpretação precisa das imagens, dificultada pela variabilidade dos exames e pela alta carga de trabalho dos radiologistas.

As altas taxas de incidência de patologias detectáveis em radiografias, como problemas na coluna lombo-sacra e anomalias nos pododáctilos, reforçam a necessidade de desenvolver pesquisas que ofereçam suporte ao diagnóstico precoce dessas condições. Nesta tese, apresentamos um sistema CAD que integra modelos de inteligência artificial preditiva e generativa para auxiliar no processo diagnóstico a partir de radiografias da coluna lombo-sacra e dos pododáctilos.

O sistema proposto visa tornar o processo de triagem de exames mais eficiente, utilizando modelos de IA preditiva para identificar e priorizar exames que necessitam de atenção especial. Além disso, empregamos modelos de IA generativa para a geração automática de laudos médicos preliminares, oferecendo uma segunda opinião que contribui para a precisão e eficiência dos diagnósticos, reduzindo a carga de trabalho dos profissionais de saúde.

Os resultados obtidos demonstram que a integração de modelos de IA preditiva e generativa em sistemas CAD pode aprimorar a precisão e a eficiência diagnóstica.Na geração automática de laudos médicos, as métricas de avaliação, como BLEU, METEOR e ROUGE-L, mostraram que os laudos gerados pelo modelo apresentam uma boa correspondência com aqueles elaborados por especialistas, especialmente para as radiografias da coluna lombo-sacra.

Embora os resultados sejam promissores, reconhecemos algumas limitações. A utilização de um conjunto de dados privado limita a possibilidade de disponibilização pública dos dados para a comunidade científica, dificultando a replicação dos resultados por outros pesquisadores. Além disso, a menor quantidade de dados disponíveis para as radiografias de pododáctilos afetou a precisão dos laudos gerados para essa região anatômica. Essa limitação ressalta a importância de conjuntos de dados amplos e diversificados para o treinamento de modelos de inteligência artificial em contextos médicos.

Em suma, este estudo contribui significativamente para o campo da inteligência

artificial aplicada à medicina, demonstrando o potencial de sistemas CAD integrados com modelos de IA preditiva e generativa para melhorar a interpretação de imagens médicas e apoiar decisões clínicas. Acreditamos que a abordagem proposta pode aprimorar a precisão e eficiência diagnóstica, reduzir a carga de trabalho dos profissionais de saúde e fornecer suporte valioso na tomada de decisão clínica, promovendo um ambiente clínico mais seguro e eficaz. Esperamos que este trabalho inspire futuras pesquisas e promova avanços adicionais na aplicação da inteligência artificial na prática médica, com o objetivo final de melhorar os resultados para os pacientes.

6.1 Trabalhos Futuros

Para aprimorar e expandir o trabalho realizado, sugerimos várias direções para futuros estudos. Uma delas é a ampliação do conjunto de dados, incorporando conjuntos mais amplos e diversificados, especialmente com maior representatividade de exames anômalos, tanto para a coluna lombo-sacra quanto para os pododáctilos. A inclusão de dados de fontes públicas ou parcerias que permitam o compartilhamento controlado dos dados pode melhorar a capacidade de generalização dos modelos e facilitar a replicação dos resultados por outros pesquisadores. Além disso, a exploração de novas arquiteturas pode contribuir para aprimorar o desempenho dos modelos. Investigando o uso de arquiteturas de redes neurais mais eficientes em termos computacionais, como MobileNet e EfficientNet para a etapa de classificação, bem como arquiteturas mais avançadas para a geração de laudos, como Transformers de última geração e modelos de linguagem pré-treinados de grande escala, podemos potencialmente alcançar melhores resultados. A experimentação com diferentes configurações de hiperparâmetros também pode contribuir para otimizar o desempenho dos modelos.

Outra área de melhoria é a incorporação de informações clínicas adicionais. Integrar dados clínicos complementares, como histórico médico, sintomas dos pacientes e resultados de outros exames, pode enriquecer a análise e a geração de laudos. Essa abordagem multimodal pode melhorar a precisão diagnóstica e fornecer insights mais profundos para os profissionais de saúde. Além disso, a avaliação em ambientes clínicos reais é fundamental. Testar e validar os modelos desenvolvidos em condições práticas permitirá avaliar seu desempenho no mundo real e obter *feedback* direto de especialistas, identificando possíveis ajustes necessários e assegurando que o sistema atenda às necessidades e expectativas dos profissionais de saúde.

Expandir a metodologia para outras modalidades de imagem médica é outra direção promissora. Adaptar a metodologia para incluir diferentes modalidades de exames clínicos, como tomografias computadorizadas, ressonâncias magnéticas ou ultrassonografias, ampliará a aplicabilidade e o impacto do sistema em diversos contextos radiológicos. Por fim, melhorias na interpretabilidade dos modelos podem aumentar a

confiança dos profissionais de saúde nas decisões tomadas pelo sistema. Investigar técnicas adicionais para aumentar a interpretabilidade dos modelos de IA permitirá que os profissionais compreendam melhor os resultados e utilizem o sistema com maior segurança.

Ao seguir essas direções, esperamos que futuros trabalhos possam aprimorar ainda mais a eficácia e eficiência dos sistemas CAD baseados em inteligência artificial, contribuindo para avanços significativos na área médica e beneficiando tanto os profissionais de saúde quanto os pacientes.

6.2 Produções Científicas

A Tabela 17 detalha os artigos diretamente relacionados ao método proposto nesta tese, destacando as contribuições específicas e os avanços na área de detecção de doenças em imagens médicas. Estes artigos evidenciam a inovação e o rigor científico da pesquisa, refletindo seu impacto significativo na comunidade acadêmica.

Tabela 17 – Produções científicas relacionada a tese em questão

Artigo	**Periódico**	**Ano**	**Tipo**	**Qualis**	**Status**
Generative Artificial Inteligence for Automated Generation of Medical Reports from Medical Images of the Lumbo-Sacral Spine using CNNs and Transformers	-	-	Periódico	-	A ser submetido
The Automated Generation of Medical Reports from Polydactyly X-ray Images Using CNNs and Transformers	Apllied Science	2024	Periódico	A2	Publicado
Deep learning approach for disease detection in lumbosacral spine radiographs using ConvNet	Computer Methods in Biomechanics and Biomedical Engineering	2023	Periódico	A4	Publicado
Detecção de Doenças em Imagens de Raios-X da Coluna Lombo-Sacra com Convnets	Simpósio Brasileiro de Computação Aplicada à Saúde	2022	Simpósio	A4	Publicado

6.3 Reconhecimento Cientıfico

Em reconhecimento ao impacto significativo da pesquisa intitulada “Detecção de Doenças em Imagens de Raios-X da Coluna Lombo-Sacra com Convnets”, destacamos que este trabalho recebeu o prêmio de Melhor Artigo Científico na área de Computação Aplicada à Saúde no Brasil, durante o SBCAS 2022. Esta distinção ressalta o prestígio e

a relevância do nosso estudo em um contexto altamente competitivo, onde concorremos com renomados pesquisadores e instituições de todo o país. A premiação é um testemunho do rigor e da inovação incorporados na nossa abordagem, e representa uma conquista notável no cenário da pesquisa em saúde e computação no Brasil.

Referências

AGARAP, A. F. Deep learning using rectified linear units (relu). **arXiv preprint arXiv:1803.08375**, 2018. Nenhuma citação no texto.

AL-JANABI, S. I. A.; LATEEF, A. A. A. Applications of deep learning approaches in speech recognition: A survey. In: SPRINGER. **Proceedings of International Conference on Computing and Communication Networks: ICCCN 2021**. [S.l.], 2022. p. 189–196. Nenhuma citação no texto.

ALHARBI, W. S.; RASHID, M. A review of deep learning applications in human genomics using next-generation sequencing data. **Human Genomics**, Springer, v. 16, n. 1, p. 26, 2022. Nenhuma citação no texto.

ANDREW, L.; SCOTT, G. Fast algorithms for convolutional neural networks. In: **2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)**. [S.l.: s.n.], 2016. p. 4013–4021. Nenhuma citação no texto.

ANNA, T.-K.; MING-HUWI, H.; CHAN-PANG, K.; MIN-JUN, F.; CHII-JEN, L.; YUNG-NIEN, S. Cobb angle measurement of spine from x-ray images using convolutional neural network. **Computational and Mathematical Methods in Medicine**, Hindawi, p. 195 – 199, 2019. ISSN 1748-670X. Disponível em: <https://doi.org/10.1155/2019/6357171>. Nenhuma citação no texto.

ANOUAR, O. I. B. P. L. C. C. D. B. L. B. Classification of coronal imbalance in adult scoliosis and spine deformity: a treatment-oriented guideline. **European Spine Journal**, 2019. Disponível em: <https://doi.org/10.1007/s00586-018-5826-3>. Nenhuma citação no texto.

ANTONIOU, A.; STORKEY, A. **Assume, Augment and Learn: Unsupervised Few-Shot Meta-Learning via Random Labels and Data Augmentation**. 2019. Nenhuma citação no texto.

BENNOUR, A.; AOUN, N. B.; KHALAF, O. I.; GHABBAN, F.; WONG, W.-K.; ALGBURI, S. Contribution to pulmonary diseases diagnostic from x-ray images using innovative deep learning models. **Heliyon**, Elsevier, v. 10, n. 9, 2024. Nenhuma citação no texto.

BERGMANN, J. N. History and mechanical control of heel spur pain. **Clinics in podiatric medicine and surgery**, Elsevier, v. 7, n. 2, p. 243–259, 1990. Nenhuma citação no texto.

BISHOP, C. M. **Pattern Recognition and Machine Learning (Information Science and Statistics)**. Berlin, Heidelberg: Springer-Verlag, 2006. ISBN 0387310738. Nenhuma citação no texto.

BOERUM, D. H. V.; SANGEORZAN, B. J. Biomechanics and pathophysiology of flat foot. **Foot and ankle clinics**, Elsevier, v. 8, n. 3, p. 419–430, 2003. Nenhuma citação no texto.

BREIMAN, L. Random forests. **Machine Learning**, v. 45, n. 1, p. 5–32, 2001. Nenhuma citação no texto.

BRIN, S.; PAGE, L. The anatomy of a large-scale hypertextual web search engine. **Computer networks and ISDN systems**, Elsevier, v. 30, n. 1-7, p. 107–117, 1998. Nenhuma citação no texto.

BROWN, T.; MANN, B.; RYDER, N.; SUBBIAH, M.; KAPLAN, J. D.; DHARIWAL, P.; NEELAKANTAN, A.; SHYAM, P.; SASTRY, G.; ASKELL, A.; AGARWAL, S.; HERBERT-VOSS, A.; KRUEGER, G.; HENIGHAN, T.; CHILD, R.; RAMESH, A.; ZIEGLER, D.; WU, J.; WINTER, C.; HESSE, C.; CHEN, M.; SIGLER, E.; LITWIN, M.; GRAY, S.; CHESS, B.; CLARK, J.; BERNER, C.; MCCANDLISH, S.; RADFORD, A.; SUTSKEVER, I.; AMODEI, D. Language models are few-shot learners. In: LAROCHELLE, H.; RANZATO, M.; HADSELL, R.; BALCAN, M.; LIN, H. (Ed.). **Advances in Neural Information Processing Systems**. Curran Associates, Inc., 2020. v. 33, p. 1877–1901. Disponível em: <https://proceedings.neurips.cc/paper_files/paper/2020/file/1457c0d6bfcb4967418bfb8ac142f64a-Paper.pdf>. Nenhuma citação no texto.

BUCHBINDER, R.; UNDERWOOD, M.; HARTVIGSEN, J.; MAHER, C. G. The lancet series call to action to reduce low value care for low back pain: an update. **Pain**, v. 161 Suppl 1, n. 1, p. S57–S64, September 2020. Nenhuma citação no texto.

BUTNARU, A. M.; IONESCU, R. T. From image to text classification: A novel approach based on clustering word embeddings. **Procedia Computer Science**, v. 112, p. 1783–1792, 2017. ISSN 1877-0509. Knowledge-Based and Intelligent Information Engineering Systems: Proceedings of the 21st International Conference, KES-20176-8 September 2017, Marseille, France. Disponível em: <https://www.sciencedirect.com/science/article/pii/S1877050917316071>. Nenhuma citação no texto.

CAIO, M.; JOHNATAN, S.; JOãO, D.; GERALDO, J.; ANSELMO, P.; JOãO, A.; SIMARA, R.; ARI, S. Diagnosis of breast cancer in images mammography through local features and invariants. **Multimedia Tools and Applications**, 2019. Nenhuma citação no texto.

CALVO-WRIGHT, M. d. M.; ÁLVARO-AFONSO, F. J.; LÓPEZ-MORAL, M.; GARCÍA-ÁLVAREZ, Y.; GARCÍA-MORALES, E.; LÁZARO-MARTÍNEZ, J. L. Is the combination of plain x-ray and probe-to-bone test useful for diagnosing diabetic foot osteomyelitis? a systematic review and meta-analysis. **Journal of Clinical Medicine**, MDPI, v. 12, n. 16, p. 5369, 2023. Nenhuma citação no texto.

CAO, Y.; CUI, L.; ZHANG, L.; YU, F.; LI, Z.; XU, Y. Mmtn: Multi-modal memory transformer network for image-report consistent medical report generation. **Proceedings of the AAAI Conference on Artificial Intelligence**, v. 37, n. 1, p. 277–285, Jun. 2023. Disponível em: <https://ojs.aaai.org/index.php/AAAI/article/view/25100>. Nenhuma citação no texto.

CAVANAGH, P.; MORAG, E.; BOULTON, A.; YOUNG, M.; DEFFNER, K.; PAMMER, S. The relationship of static foot structure to dynamic foot function. **Journal of Biomechanics**, v. 30, n. 3, p. 243–250, 1997. ISSN 0021-9290. Disponível em: <https://www.sciencedirect.com/science/article/pii/S0021929096001364>. Nenhuma citação no texto.

CHEN, T.; GUESTRIN, C. Xgboost. **Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining**, ACM, Aug 2016. Disponível em: <http://dx.doi.org/10.1145/2939672.2939785>. Nenhuma citação no texto.

CHERUKURI, M.; STANLEY, R.; LONG, R.; ANTANI, S.; THOMA, G. Anterior osteophyte discrimination in lumbar vertebrae using size-invariant features. **Computerized Medical Imaging and Graphics**, v. 28, n. 1, p. 99 – 108, 2004. ISSN 0895-6111. Disponível em: <http://www.sciencedirect.com/science/article/pii/S0895611103000752>. Nenhuma citação no texto.

CIEZA, A.; CAUSEY, K.; KAMENOV, K.; HANSON, S. W.; CHATTERJI, S.; VOS, T. Global estimates of the need for rehabilitation based on the global burden of disease study 2019: a systematic analysis for the global burden of disease study 2019. **The Lancet**, Elsevier, v. 396, n. 10267, p. 2006–2017, December 2020. Open Access. Disponível em: <https://doi.org/10.1016/S0140-6736(20)32340-0>. Nenhuma citação no texto.

COHEN, J. Weighted kappa: Nominal scale agreement provision for scaled disagreement or partial credit. **Psychological bulletin**, American Psychological Association, v. 70, n. 4, p. 213, 1968. Nenhuma citação no texto.

COLLOBERT, R.; WESTON, J.; BOTTOU, L.; KARLEN, M.; KAVUKCUOGLU, K.; KUKSA, P. Natural language processing (almost) from scratch. **Journal of machine learning research**, v. 12, p. 2493–2537, 2011. Nenhuma citação no texto.

CONGALTON, R. G. Accuracy assessment: A user's perspective. **Photogrammetric Engineering and Remote Sensing**, v. 52, p. 397–399, 1986. Nenhuma citação no texto.

CONNOR, K. T. M. S. A survey on Image Data Augmentation for Deep Learning. **Journal of Big Data**, 07 2019. ISSN 2196-1115. Nenhuma citação no texto.

CORTES, C.; VAPNIK, V. Support-vector networks. **Machine Learning**, v. 20, n. 3, p. 273–297, Sep 1995. ISSN 1573-0565. Nenhuma citação no texto.

CUBUK, E. D.; ZOPH, B.; MANÉ, D.; VASUDEVAN, V.; LE, Q. V. Autoaugment: Learning augmentation policies from data. **CoRR**, abs/1805.09501, 2018. Disponível em: <http://arxiv.org/abs/1805.09501>. Nenhuma citação no texto.

CUBUK, E. D.; ZOPH, B.; SHLENS, J.; LE, Q. V. Randaugment: Practical data augmentation with no separate search. **CoRR**, abs/1909.13719, 2019. Disponível em: <http://arxiv.org/abs/1909.13719>. Nenhuma citação no texto.

DABOUEI, A.; SOLEYMANI, S.; TAHERKHANI, F.; NASRABADI, N. M. **SuperMix: Supervising the Mixing Data Augmentation**. 2020. Nenhuma citação no texto.

DENKOWSKI, M.; LAVIE, A. Meteor universal: Language specific translation evaluation for any target language. In: BOJAR, O.; BUCK, C.; FEDERMANN, C.; HADDOW, B.; KOEHN, P.; MONZ, C.; POST, M.; SPECIA, L. (Ed.). **Proceedings of the Ninth Workshop on Statistical Machine Translation**. Baltimore, Maryland, USA: Association for Computational Linguistics, 2014. p. 376–380. Disponível em: <https://aclanthology.org/W14-3348>. Nenhuma citação no texto.

DESCHAMPS, K.; BIRCH, I.; DESLOOVERE, K.; MATRICALI, G. A. The impact of hallux valgus on foot kinematics: A cross-sectional, comparative study. **Gait Posture**, v. 32, n. 1, p. 102–106, 2010. ISSN 0966-6362. Disponível em: <https://www.sciencedirect.com/science/article/pii/S0966636210001013>. Nenhuma citação no texto.

DEVLIN, J.; CHANG, M.-W.; LEE, K.; TOUTANOVA, K. Bert: Pre-training of deep bidirectional transformers for language understanding. **arXiv preprint arXiv:1810.04805**, 2018. Nenhuma citação no texto.

DRAKE, R.; VOGL, W.; MITCHELL, A. **Gray's Anatomy for Students**. Elsevier, 2019. (Gray's Anatomy Series). ISBN 9780323393041. Disponível em: <https://books.google.com.br/books?id=7oy4vwEACAAJ>. Nenhuma citação no texto.

DUARTE, R.; NEMMEN, R.; NAVARRO, J. P. Black hole weather forecasting with deep learning: a pilot study. **Monthly Notices of the Royal Astronomical Society**, v. 512, n. 4, p. 5848–5861, 03 2022. ISSN 0035-8711. Disponível em: <https://doi.org/10.1093/mnras/stac665>. Nenhuma citação no texto.

ESMAIL, K. M.; EL-DIN, H. E.; A., S. M. Cascaded deep learning classifiers for computer-aided diagnosis of covid-19 and pneumonia diseases in x-ray scans. **Complex & Intelligent Systems**, 2020. ISSN 2198-6053. Disponível em: <https://doi.org/10.1007/s40747-020-00199-4>. Nenhuma citação no texto.

FLEISS, J. L.; COHEN, J. The equivalence of weighted kappa and the intraclass correlation coefficient as measures of reliability. **Educational and Psychological Measurement**, v. 33, n. 3, p. 613–619, 1973. Disponível em: <https://doi.org/10.1177/001316447303300309>. Nenhuma citação no texto.

GARDNER, M. W.; DORLING, S. R. **Artificial Neural Networks (The Multilayer Perceptron) A Review of Applications in the Atmospheric Sciences**. [S.l.], 1998. v. 32, 627-263 p. Nenhuma citação no texto.

GEBO, D. L. Foot morphology and locomotor adaptation in eocene primates. **Folia Primatologica**, Brill, Leiden, The Netherlands, v. 50, n. 1-2, p. 3 – 41, 1988. Disponível em: <https://brill.com/view/journals/ijfp/50/1-2/article-p3_2.xml>. Nenhuma citação no texto.

GEIRHOS, R.; JACOBSEN, J.-H.; MICHAELIS, C.; ZEMEL, R.; BRENDEL, W.; BETHGE, M.; WICHMANN, F. A. **Shortcut Learning in Deep Neural Networks**. 2020. Nenhuma citação no texto.

GOLLERY, M. Bioinformatics: sequence and genome analysis. **Clinical Chemistry**, Oxford University Press, v. 51, n. 11, p. 2219, 2005. Nenhuma citação no texto.

GONG, Y.; COSMA, G.; FANG, H. On the limitations of visual-semantic embedding networks for image-to-text information retrieval. **Journal of Imaging**, v. 7, n. 8, 2021. ISSN 2313-433X. Disponível em: <https://www.mdpi.com/2313-433X/7/8/125>. Nenhuma citação no texto.

GONZALEZ, R. E. W. R. C. **Processamento digital de imagens.** 3. ed. [S.l.]: Pearson, 2011. ISSN 8576054019. Nenhuma citação no texto.

GOODFELLOW, I.; BENGIO, Y.; COURVILLE, A. **Deep Learning**. [S.l.]: MIT Press, 2016. 180-184. p. <http://www.deeplearningbook.org>. Nenhuma citação no texto.

GOODFELLOW YOSHUA BENGIO, A. C. I. **Deep Learning**. [S.l.]: MIT Press, 2016. ISBN 9780262035613. Nenhuma citação no texto.

GRUSHKY, A. D.; IM, S. J.; STEENBURG, S. D.; CHONG, S. Traumatic injuries of the foot and ankle. In: ELSEVIER. **Seminars in roentgenology**. [S.l.], 2021. v. 56, n. 1, p. 47–69. Nenhuma citação no texto.

GU, J.; LI, C.; LIANG, Y.; SHI, Z.; SONG, Z. Exploring the frontiers of softmax: Provable optimization, applications in diffusion model, and beyond. **arXiv preprint arXiv:2405.03251**, 2024. Nenhuma citação no texto.

HAND, R. J. T. D. J. A simple generalisation of the area under the roc curve for multiple class classification problems. **Machine Learning**, v. 45, p. 171–186, 2001. Nenhuma citação no texto.

HAWES, M. R.; SOVAK, D. Quantitative morphology of the human foot in a north american population. **Ergonomics**, Taylor & Francis, v. 37, n. 7, p. 1213–1226, 1994. PMID: 8050406. Disponível em: <https://doi.org/10.1080/00140139408964899>. Nenhuma citação no texto.

HAYKIN, S. **Redes Neurais: Princípios e Prática**. Artmed, 2007. ISBN 9788577800865. Disponível em: <https://books.google.com.br/books?id=bhMwDwAAQBAJ>. Nenhuma citação no texto.

HE, K.; ZHANG, X.; REN, S.; SUN, J. Deep residual learning for image recognition. **CoRR**, abs/1512.03385, 2015. Nenhuma citação no texto.

HE, K.; ZHANG, X.; REN, S.; SUN, J. **Deep Residual Learning for Image Recognition**. 2015. Nenhuma citação no texto.

HILL, R.; HEALY, B.; HOLLOWAY, L.; KUNCIC, Z.; THWAITES, D.; BALDOCK, C. Advances in kilovoltage x-ray beam dosimetry. **Physics in Medicine and Biology**, IOP Publishing, v. 59, n. 6, p. R183–R231, feb 2014. Disponível em: <https://doi.org/10.1088/0031-9155/59/6/r183>. Nenhuma citação no texto.

HUANG, X.; YAN, F.; XU, W.; LI, M. Multi-attention and incorporating background information model for chest x-ray image report generation. **IEEE Access**, v. 7, p. 154808–154817, 2019. Nenhuma citação no texto.

ISLAM, S.; ELMEKKI, H.; ELSEBAI, A.; BENTAHAR, J.; DRAWEL, N.; RJOUB, G.; PEDRYCZ, W. A comprehensive survey on applications of transformers for deep learning tasks. **Expert Systems with Applications**, v. 241, p. 122666, 2024. ISSN 0957-4174. Disponível em: <https://www.sciencedirect.com/science/article/pii/S0957417423031688>. Nenhuma citação no texto.

JOHN., L. **Modern Cosmology & Philosophy**. [S.l.]: University of Michigan: Prometheus Books, 1998. ISSN 978-1573922500. Nenhuma citação no texto.

KALPIĆ, D.; HLUPIĆ, N.; LOVRIĆ, M. Student's t-tests. In: ____. **International Encyclopedia of Statistical Science**. Berlin, Heidelberg: Springer Berlin Heidelberg, 2011. p. 1559–1563. ISBN 978-3-642-04898-2. Disponível em: <https://doi.org/10.1007/978-3-642-04898-2_641>. Nenhuma citação no texto.

KARPAGAVALLI, S.; CHANDRA, E. A review on automatic speech recognition architecture and approaches. **International Journal of Signal Processing, Image Processing and Pattern Recognition**, v. 9, n. 4, p. 393–404, 2016. Nenhuma citação no texto.

KASHIRI, N.; ABATE, A.; ABRAM, S. J.; ALBU-SCHAFFER, A.; CLARY, P. J.; DALEY, M.; FARAJI, S.; FURNEMONT, R.; GARABINI, M.; GEYER, H. et al. An overview on principles for energy efficient robot locomotion. **Frontiers in Robotics and AI**, Frontiers Media SA, v. 5, p. 129, 2018. Nenhuma citação no texto.

KETCHAM ROGER W. LOWE, J. W. W. D. J. **IMAGE ENHANCEMENT TECHNIQUES FOR COCKPIT DISPLAYS**. [S.l.], 1974. Nenhuma citação no texto.

KIM, H.-G.; LEE, K. M.; KIM, E. J.; LEE, J. S. Improvement diagnostic accuracy of sinusitis recognition in paranasal sinus x-ray using multiple deep learning models. **Quantitative imaging in medicine and surgery**, AME Publications, v. 9, n. 6, p. 942, 2019. Nenhuma citação no texto.

KIM, J.-H.; LEE, S.-E.; JUNG, H.-S.; SHIM, B.-S.; HOU, J.-U.; KWON, Y.-S. Development and validation of deep learning-based algorithms for predicting lumbar herniated nucleus pulposus using lumbar x-rays. **Journal of Personalized Medicine**, v. 12, n. 5, 2022. ISSN 2075-4426. Disponível em: <https://www.mdpi.com/2075-4426/12/5/767>. Nenhuma citação no texto.

KING, G.; ZENG, L. Logistic regression in rare events data. **Political Analysis**, v. 9, p. 137 – 163, 2001. Nenhuma citação no texto.

KINGMA, D. P.; BA, J. **Adam: A Method for Stochastic Optimization**. 2017. Nenhuma citação no texto.

KOHA, R. A study of crossvalidation and bootstrap for accuracy estimation and model selectivv. **Appears in the International Joint Conference on Articial Intelligence IJCAI**, 1995. Nenhuma citação no texto.

KOHAVI, R. Glossary of terms. **Machine learning**, v. 30, p. 271–274, 1998. Nenhuma citação no texto.

KONG, J.-W.; OH, B.-D.; KIM, C.; KIM, Y.-S. Sequential brain ct image captioning based on the pre-trained classifiers and a language model. **Applied Sciences**, v. 14, n. 3, 2024. ISSN 2076-3417. Disponível em: <https://www.mdpi.com/2076-3417/14/3/1193>. Nenhuma citação no texto.

KOREZ, R.; PUTZIER, M.; VRTOVEC, T. A deep learning tool for fully automated measurements of sagittal spinopelvic balance from x-ray images: performance evaluation. **European Spine Journal**, v. 29, n. 9, p. 2295–2305, 09 2020. Disponível em: <https://doi.org/10.1007/s00586-020-06406-7>. Nenhuma citação no texto.

KOUGIA, V.; PAVLOPOULOS, J.; PAPAPETROU, P.; GORDON, M. RTEX: A novel framework for ranking, tagging, and explanatory diagnostic captioning of radiography exams. **Journal of the American Medical Informatics Association**, v. 28, n. 8, p. 1651–1659, 04 2021. ISSN 1527-974X. Disponível em: <https://doi.org/10.1093/jamia/ocab046>. Nenhuma citação no texto.

KRIZHEVSKY, A.; SUTSKEVER, I.; HINTON, G. E. Imagenet classification with deep convolutional neural networks. **Communications of the ACM**, AcM New York, NY, USA, v. 60, n. 6, p. 84–90, 2017. Nenhuma citação no texto.

KRZYWINSKI, M.; ALTMAN, N. Points of significance: Significance, p values and t-tests. **Nature methods**, v. 10, n. 11, 2013. Nenhuma citação no texto.

LECUN, Y.; L, B.; Y, B.; P, H. Gradient-based learning applied to document recognition. **Proceedings of the IEEE**, v. 86, n. 11, p. 2278–2324, 1998. Nenhuma citação no texto.

LECUN YOSHUA BENGIO, G. H. Y. Deep learning. **Nature**, v. 521, 2015. Disponível em: <https://doi.org/10.1038/nature14539>. Nenhuma citação no texto.

LEE, C.; JANG, J.; LEE, S.; KIM, Y. S.; JO, H. J.; KIM, Y. Classification of femur fracture in pelvic x-ray images using meta-learned deep neural network. **Scientific reports**, Nature Publishing Group UK London, v. 10, n. 1, p. 13694, 2020. Nenhuma citação no texto.

LEE, C.-Y.; GALLAGHER, P. W.; TU, Z. Generalizing pooling functions in convolutional neural networks: Mixed, gated, and tree. In: GRETTON, A.; ROBERT, C. C. (Ed.). **Proceedings of the 19th International Conference on Artificial Intelligence and Statistics**. Cadiz, Spain: PMLR, 2016. (Proceedings of Machine Learning Research, v. 51), p. 464–472. Disponível em: <http://proceedings.mlr.press/v51/lee16a.html>. Nenhuma citação no texto.

LEE, H. S.; KANG, J.; KIM, S. E.; KIM, J. H.; CHO, B.-J. Estimating infant age from skull x-ray images using deep learning. **Scientific Reports**, Nature Publishing Group UK London, v. 14, n. 1, p. 16600, 2024. Nenhuma citação no texto.

LEE, S.; CHOE, E.; KANG, H.; YOON, J.; KIM, H. The exploration of feature extraction and machine learning for predicting bone density from simple spine x-ray images in a korean population. **Skeletal Radiology**, v. 49, 11 2019. Nenhuma citação no texto.

LI, H.; ZOU, L.; KOWAH, J. A. H.; HE, D.; LIU, Z.; DING, X.; WEN, H.; WANG, L.; YUAN, M.; LIU, X. A compact review of progress and prospects of deep learning in drug discovery. **Journal of Molecular Modeling**, Springer, v. 29, n. 4, p. 117, 2023. Nenhuma citação no texto.

LI, Z.; HOIEM, D. Learning without forgetting. In: LEIBE, B.; MATAS, J.; SEBE, N.; WELLING, M. (Ed.). **Computer Vision – ECCV 2016**. Cham: Springer International Publishing, 2016. p. 614–629. ISBN 978-3-319-46493-0. Nenhuma citação no texto.

LIN, C.-Y. ROUGE: A package for automatic evaluation of summaries. In: **Text Summarization Branches Out**. Barcelona, Spain: Association for Computational Linguistics, 2004. p. 74–81. Disponível em: <https://aclanthology.org/W04-1013>. Nenhuma citação no texto.

LIU, B.; WANG, X.; DIXIT, M.; KWITT, R.; VASCONCELOS, N. Feature space transfer for data augmentation. In: **Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)**. [S.l.: s.n.], 2018. Nenhuma citação no texto.

LIU, S.; GUO, C.; ZHAO, Y.; ZHANG, C.; YUE, L.; YAO, R.; LAN, Q.; ZHOU, X.; ZHAO, B.; WU, J.; LI, W.; XU, N. A machine learning based quantification system for automated diagnosis of lumbar spondylolisthesis on spinal x-rays. **Heliyon**, Elsevier, v. 10, n. 17, p. e37418, Sep 2024. ISSN 2405-8440. Disponível em: <https://doi.org/10.1016/j.heliyon.2024.e37418>. Nenhuma citação no texto.

LIU, Y.; OTT, M.; GOYAL, N.; DU, J.; JOSHI, M.; CHEN, D.; LEVY, O.; LEWIS, M.; ZETTLEMOYER, L.; STOYANOV, V. **RoBERTa: A Robustly Optimized BERT Pretraining Approach**. 2019. Nenhuma citação no texto.

LONG, S.; HE, X.; YAO, C. Scene text detection and recognition: The deep learning era. **International Journal of Computer Vision**, Springer, v. 129, n. 1, p. 161–184, 2021. Nenhuma citação no texto.

MANAKITSA, N.; MARASLIDIS, G. S.; MOYSIS, L.; FRAGULIS, G. F. A review of machine learning and deep learning for object detection, semantic segmentation, and human action recognition in machine and robotic vision. **Technologies**, v. 12, n. 2, 2024. ISSN 2227-7080. Disponível em: <https://www.mdpi.com/2227-7080/12/2/15>. Nenhuma citação no texto.

MATTHEWS, J. The developmental anatomy of the foot. **The Foot**, v. 8, n. 1, p. 17–25, 1998. ISSN 0958-2592. Disponível em: <https://www.sciencedirect.com/science/article/pii/S0958259298900153>. Nenhuma citação no texto.

MAULUD, D.; ABDULAZEEZ, A. M. A review on linear regression comprehensive in machine learning. **Journal of Applied Science and Technology Trends**, v. 1, n. 2, p. 140–147, 2020. Nenhuma citação no texto.

MEANING of overfitting in English: overfitting. Oxford University Press, 1933. Disponível em: <https://www.lexico.com/definition/overfitting>. Nenhuma citação no texto.

MINSKY, S. P. M. **Perceptrons - Expanded Edition**. [S.l.]: The MIT Press, 1988. ISBN 0-262-63111-3. Nenhuma citação no texto.

MOHSAN, M. M.; AKRAM, M. U.; RASOOL, G.; ALGHAMDI, N. S.; BAQAI, M. A. A.; ABBAS, M. Vision transformer and language model based radiology report generation. **IEEE Access**, v. 11, p. 1814–1824, 2023. Nenhuma citação no texto.

NAGUIB, S. M.; HAMZA, H. M.; HOSNY, K. M.; SALEH, M. K.; KASSEM, M. A. Classification of cervical spine fracture and dislocation using refined pre-trained deep model and saliency map. **Diagnostics**, MDPI, v. 13, n. 7, p. 1273, 2023. Nenhuma citação no texto.

NR, S. M. P. Detection of Pneumonia in chest X-ray images. **Journal of X-ray science and technology**, 04 2011. ISSN 1095-9114. Nenhuma citação no texto.

O.A., Y. A. D. A. P. Spondyloarthrosis: pathogenesis, clinic, diagnosis and treatment (literature review and own experience). **Journal of Clinical Practice**, v. 10, n. 4, p. 61–73, 2019. Nenhuma citação no texto.

OLSON, D. D. D. L. **Advanced Data Mining Techniques**. 1. ed. [S.l.]: Springer-Verlag Berlin Heidelberg, 2008. XII, 180 p. ISSN 978-3-540-76916-30. Nenhuma citação no texto.

OTSU, N. A Threshold Selection Method from Gray-level Histograms. **IEEE Transactions on Systems, Man and Cybernetics**, v. 9, n. 1, p. 62–66, 1979. Nenhuma citação no texto.

PAPINENI, K.; ROUKOS, S.; WARD, T.; ZHU, W.-J. Bleu: a method for automatic evaluation of machine translation. In: **Proceedings of the 40th Annual Meeting on Association for Computational Linguistics**. USA: Association for Computational Linguistics, 2002. (ACL '02), p. 311–318. Disponível em: <https://doi.org/10.3115/1073083.1073135>. Nenhuma citação no texto.

PARAS, L. Deep Convolutional Neural Networks for Endotracheal Tube Position and X-ray Image Classification: Challenges and Opportunities. **Journal of Digital Imaging**, 08 2017. ISSN 1618-727X. Nenhuma citação no texto.

PAVLOPOULOS, J.; KOUGIA, V.; ANDROUTSOPOULOS, I.; PAPAMICHAIL, D. Diagnostic captioning: a survey. **Knowledge and Information Systems**, v. 64, n. 7, p. 1691–1722, 2022. ISSN 0219-3116. Disponível em: <https://doi.org/10.1007/s10115-022-01684-7>. Nenhuma citação no texto.

PEDRINI, H.; SCHWARTZ, W. R. **Análise de imagens digitais: princípios, algoritmos e aplicações**. [S.l.]: Thomson Learning, 2008. Nenhuma citação no texto.

PENSEC, V. D.; SARAUX, A.; BERTHELOT, J. M.; ALAPETITE, S.; JOUSSE, S.; CHALES, G.; THOREL, J. B.; HOANG, S.; NOUY-TROLLE, I.; MARTIN, A. et al. Ability of foot radiographs to predict rheumatoid arthritis in patients with early arthritis. **The Journal of Rheumatology**, The Journal of Rheumatology, v. 31, n. 1, p. 66–70, 2004. Nenhuma citação no texto.

PIZER, S. M.; JOHNSTON, R. E.; ERICKSEN, J. P.; YANKASKAS, B. C.; MULLER, K. E. Contrast-limited adaptive histogram equalization: Speed and effectiveness. **IEEE**, 1990. Nenhuma citação no texto.

POPESCU, M.-C.; BALAS, V. E.; PERESCU-POPESCU, L.; MASTORAKIS, N. Multilayer perceptron and neural networks. **WSEAS Transactions on Circuits and Systems**, World Scientific and Engineering Academy and Society (WSEAS), Stevens Point, Wisconsin, USA, v. 8, n. 7, p. 579–588, jul. 2009. ISSN 1109-2734. Nenhuma citação no texto.

PURVES, D.; AGOSTINHO, G. J.; FITZPATRICK, D.; HALL, W. C.; AMANTIA, A.-S. L.; MCNAMARA, J. O.; WHITE, L. E. **Neurociências**. [S.l.]: ARTMED EDITORA SA, 2010. ISBN 978-0-87893-697-7. Nenhuma citação no texto.

RANJAN, A.; BEHERA, V. N. J.; REZA, M. Ocr using computer vision and machine learning. **Machine Learning Algorithms for Industrial Applications**, Springer, p. 83–105, 2021. Nenhuma citação no texto.

RAZAVIAN, A. S.; AZIZPOUR, H.; SULLIVAN, J.; CARLSSON, S. **CNN Features off-the-shelf: an Astounding Baseline for Recognition**. 2014. Nenhuma citação no texto.

REUMATOLOGIA, S. B. de. **Radiografias Simples**. 2011. <https://www.reumatologia.org.br/orientacoes-ao-paciente/radiografias-simples/>, (Acessado 01 Fevereiro de 2021). Nenhuma citação no texto.

RIJSBERGEN, C. V. **Information Retrieval**. [S.l.]: Butterworth-Heinemann, 1979. v. 2. ISBN 0408709294. Nenhuma citação no texto.

RODDY, E.; MENZ, H. B. Foot osteoarthritis: latest evidence and developments. **Therapeutic Advances in Musculoskeletal Disease**, SAGE Publications Sage UK: London, England, v. 10, n. 4, p. 91–103, 2018. Nenhuma citação no texto.

RONNEBERGER, O.; P.FISCHER; BROX, T. U-net: Convolutional networks for biomedical image segmentation. In: **Medical Image Computing and Computer-Assisted Intervention (MICCAI)**. [S.l.]: Springer, 2015. (LNCS, v. 9351), p. 234–241. Nenhuma citação no texto.

ROSENBLATT, F. **The Perceptron: A Probabilistic Model for Information Storage and Organization in the Brain**. [S.l.], 1958. Nenhuma citação no texto.

ROUX, N. L.; BENGIO, Y.; FITZGIBBON, A. Improving first and second-order methods by modeling uncertainty. 2011. Nenhuma citação no texto.

RUMELHART, D.; HINTON, G.; WILLIAMS, R. **Learning representations by back-propagating errors**. [S.l.], 1986. Nenhuma citação no texto.

RUSSAKOVSKY, O.; DENG, J.; SU, H.; KRAUSE, J.; SATHEESH, S.; MA, S.; HUANG, Z.; KARPATHY, A.; KHOSLA, A.; BERNSTEIN, M.; BERG, A. C.; FEI-FEI, L. ImageNet Large Scale Visual Recognition Challenge. **International Journal of Computer Vision (IJCV)**, v. 115, n. 3, p. 211–252, 2015. Nenhuma citação no texto.

SAAD, O.; DARWISH, A.; FARAJ, R. A survey of machine learning techniques for spam filtering. **International Journal of Computer Science and Network Security (IJCSNS)**, International Journal of Computer Science and Network Security, v. 12, n. 2, p. 66, 2012. Nenhuma citação no texto.

SALEEM, M.; FARID, M. S.; SALEEM, S.; KHAN, M. H. X-ray image analysis for automated knee osteoarthritis detection. **Signal, Image and Video Processing**, Springer, v. 14, n. 6, p. 1079–1087, 2020. Nenhuma citação no texto.

SALTZMAN, C. L.; NAWOCZENSKI, D. A. Complexities of foot architecture as a base of support. **Journal of Orthopaedic & Sports Physical Therapy**, v. 21, n. 6, p. 354–360, 1995. Disponível em: <https://www.jospt.org/doi/10.2519/jospt.1995.21.6.354>. Nenhuma citação no texto.

SAMUEL, A. L. Some studies in machine learning using the game of checkers. **IBM Journal of Research and Development**, v. 3, n. 3, p. 210–229, 1959. Nenhuma citação no texto.

SANGNARK, S.; RATTANACHAISIT, P.; PATCHARATRAKUL, T.; VATEEKUL, P. Explainable multi-modal deep learning with cross-modal attention for diagnosis of dyssynergic defecation using abdominal x-ray images and symptom questionnaire. **IEEE Access**, IEEE, 2024. Nenhuma citação no texto.

SARAIVA1, A. A.; FERREIRA, N. M. F.; SOUSA, L. L. de; COSTA, N. J. C.; SOUSA, J. V. M.; SANTOS, D. B. S.; VALENTE, A.; SOARES, S. Classification of images of childhood pneumonia using convolutional neural networks. **In Proceedings of the 12th International Joint Conference on Biomedical Engineering Systems and Technologies (BIOSTEC)**, 2019. Nenhuma citação no texto.

SARKER, I. H. Machine learning: Algorithms, real-world applications and research directions. **SN computer science**, Springer, v. 2, n. 3, p. 160, 2021. Nenhuma citação no texto.

SCHEIBEL, P. C.; MATHEUS, P. D.; ALBINO, C. C.; RAMOS, A. L. Correlação entre a densidade óssea mandibular, femural, lombar e cervical. **Revista Dental Press de Ortodontia e Ortopedia Facial**, SciELO Brasil, v. 14, p. 111–122, 2009. Nenhuma citação no texto.

SCHERER ANDREAS MULLER, S. B. D. Evaluation of pooling operations in convolutional architectures for object recognition. **20th International Conference on Artificial Neural Networks (ICANN)**, 2010. Nenhuma citação no texto.

SCHWARTZ, J. T. B.; CHO, B. H. B.; TANG, P. M.; SCHEFFLEIN, J. M.; ARVIND, V. B.; KIM, J. S. M.; DOSHI, A. H. M.; CHO, S. K. M. Deep learning automates measurement of spinopelvic parameters on lateral lumbar radiographs. **SPINE**, v. 46, n. 12, p. E671–E678, June 2021. Nenhuma citação no texto.

SELVARAJU, R. R.; DAS, A.; VEDANTAM, R.; COGSWELL, M.; PARIKH, D.; BATRA, D. **Grad-CAM: Why did you say that?** 2016. Nenhuma citação no texto.

SEMA, A. S. C. A review on lung boundary detection in chest x-rays. **International Journal of Computer Assisted Radiology and Surgery**, v. 14, n. 6, p. R183–R231, 2019. Disponível em: <https://doi.org/10.1007/s11548-019-01917-1>. Nenhuma citação no texto.

SERRA-BURRIEL, M.; AMES, C. Machine learning-based clustering analysis: foundational concepts, methods, and applications. In: SPRINGER. **Machine Learning in Clinical Neuroscience: Foundations and Applications**. [S.l.], 2022. p. 91–100. Nenhuma citação no texto.

SHAIK, N. S.; CHERUKURI, T. K. Gated contextual transformer network for multi-modal retinal image clinical description generation. **Image and Vision Computing**, v. 143, p. 104946, 2024. ISSN 0262-8856. Disponível em: <https://www.sciencedirect.com/science/article/pii/S0262885624000490>. Nenhuma citação no texto.

SHAMSHAD, F.; KHAN, S.; ZAMIR, S. W.; KHAN, M. H.; HAYAT, M.; KHAN, F. S.; FU, H. Transformers in medical imaging: A survey. **Medical Image Analysis**, Elsevier, p. 102802, 2023. Nenhuma citação no texto.

SHEN, L.; GAO, C.; HU, S.; KANG, D.; ZHANG, Z.; XIA, D.; XU, Y.; XIANG, S.; ZHU, Q.; XU, G. et al. Using artificial intelligence to diagnose osteoporotic vertebral fractures on plain radiographs. **Journal of Bone and Mineral Research**, John Wiley & Sons, Inc. Hoboken, USA, v. 38, n. 9, p. 1278–1287, 2023. Nenhuma citação no texto.

SHRIVASTAVA, N.; BHARTI, J. Empirical analysis of image segmentation techniques. In: UNAL, A.; NAYAK, M.; MISHRA, D. K.; SINGH, D.; JOSHI, A. (Ed.). **Smart Trends in Information Technology and Computer Communications**. Singapore: Springer Singapore, 2016. p. 143–150. ISBN 978-981-10-3433-6. Nenhuma citação no texto.

SIMON, P. **Too big to ignore: the business case for big data**. [S.l.]: John Wiley & Sons, 2013. v. 72. Nenhuma citação no texto.

SIMONYAN, K.; ZISSERMAN, A. **Very Deep Convolutional Networks for Large-Scale Image Recognition**. 2014. Nenhuma citação no texto.

SIMONYAN, K.; ZISSERMAN, A. Very deep convolutional networks for large-scale image recognition. In: **International Conference on Learning Representations**. [S.l.: s.n.], 2015. Nenhuma citação no texto.

SINGH, S. A.; MAJUMDER, S. Chapter one - short and noisy electrocardiogram classification based on deep learning. In: DAS, H.; PRADHAN, C.; DEY, N. (Ed.). **Deep Learning for Data Analytics**. [S.l.]: Academic Press, 2020. p. 1–19. ISBN 978-0-12-819764-6. Nenhuma citação no texto.

SINGH, S. A.; MEITEI, T. G.; MAJUMDER, S. 6 - short pcg classification based on deep learning. In: AGARWAL, B.; BALAS, V. E.; JAIN, L. C.; POONIA, R. C.; MANISHA (Ed.). **Deep Learning Techniques for Biomedical and Health Informatics**. Academic Press, 2020. p. 141–164. ISBN 978-0-12-819061-6. Disponível em: <https://www.sciencedirect.com/science/article/pii/B9780128190616000069>. Nenhuma citação no texto.

SIRIPATTANADILOK, W.; SIRIBORVORNRATANAKUL, T. Recognition of partially occluded soft-shell mud crabs using faster r-cnn and grad-cam. **Aquaculture International**, Springer, v. 32, n. 3, p. 2977–2997, 2024. Nenhuma citação no texto.

STEVENS, B. An analysis of the structure and brevity of preliminary clinical evaluations describing traumatic abnormalities on extremity x-ray images. **Radiography**, Elsevier, v. 26, n. 4, p. 302–307, 2020. Nenhuma citação no texto.

SZEGEDY, C.; VANHOUCKE, V.; IOFFE, S.; SHLENS, J.; WOJNA, Z. **Rethinking the Inception Architecture for Computer Vision**. 2015. Nenhuma citação no texto.

TAJBAKHSH, N.; SHIN, J.; GURUDU, S.; HURST, R.; KENDALL, C.; GOTWAY, M.; LIANG, J. Convolutional neural networks for medical image analysis: Full training or fine tuning? **IEEE Transactions on Medical Imaging**, v. 35, n. 5, p. 1299–1312, 2016. Nenhuma citação no texto.

TELEA, A. An image inpainting technique based on the fast marching method. **Journal of Graphics Tools**, Taylor & Francis, v. 9, n. 1, p. 23–34, 2004. Nenhuma citação no texto.

THARWAT, A. Classification assessment methods. **Applied Computing and Informatics**, August 2020. ISSN 2634-1964. Nenhuma citação no texto.

TOMASSONI, D.; TRAINI, E.; AMENTA, F. Gender and age related differences in foot morphology. **Maturitas**, v. 79, n. 4, p. 421–427, 2014. ISSN 0378-5122. Disponível em: <https://www.sciencedirect.com/science/article/pii/S0378512214002606>. Nenhuma citação no texto.

TRAN, N.-T.; TRAN, V.-H.; NGUYEN, N.-B.; NGUYEN, T.-K.; CHEUNG, N.-M. On data augmentation for gan training. **IEEE Transactions on Image Processing**, Institute of Electrical and Electronics Engineers (IEEE), v. 30, p. 1882–1897, 2021. ISSN 1941-0042. Disponível em: <http://dx.doi.org/10.1109/TIP.2021.3049346>. Nenhuma citação no texto.

TREVOR, C.; ROSAIRE, M.; DRISCOLL, M. Vacuum curette lumbar discectomy mechanics for use in spine surgical training simulators. **Scientific Reports**, v. 12, 08 2022. Nenhuma citação no texto.

TROJIAN, T.; TUCKER, A. K. Plantar fasciitis. **American family physician**, v. 99, n. 12, p. 744–750, 2019. Nenhuma citação no texto.

TSANIYA, H.; FATICHAH, C.; SUCIATI, N. Automatic radiology report generator using transformer with contrast-based image enhancement. **IEEE Access**, v. 12, p. 25429–25442, 2024. Nenhuma citação no texto.

VASWANI, A.; SHAZEER, N.; PARMAR, N.; USZKOREIT, J.; JONES, L.; GOMEZ, A. N.; KAISER, L. u.; POLOSUKHIN, I. Attention is all you need. In: GUYON, I.; LUXBURG, U. V.; BENGIO, S.; WALLACH, H.; FERGUS, R.; VISHWANATHAN, S.; GARNETT, R. (Ed.). **Advances in Neural Information Processing Systems**. Curran Associates, Inc., 2017. v. 30. Disponível em: <https://proceedings.neurips.cc/paper_files/paper/2017/file/3f5ee243547dee91fbd053c1c4a845aa-Paper.pdf>. Nenhuma citação no texto.

VERONEZI, C. C. D.; SIMAMUES, P. W. T. d. A.; SANTOS, R. L. d.; ROCHA, E. L. d.; MELAPOUNDSO, S.; MATTOS, M. C. d.; CECHINEL, C. Computational analysis based on artificial neural networks for aiding in diagnosing osteoarthritis of the lumbar spine. **Revista Brasileira de Ortopedia**, scielo, v. 46, p. 195 – 199, 04 2011. ISSN 0102-3616. Nenhuma citação no texto.

VEXELS. **Ilustração de Neurônio**. [S.l.], 2017. Disponível em: <https://br.vexels.com/png-svg/previsualizar/145055/ilustracao-de-neuronio>. Nenhuma citação no texto.

VIEIRA, P.; SOUSA, O.; MAGALHES, D.; RABLO, R.; SILVA, R. Detecting pulmonary diseases using deep features in x-ray images. **Pattern Recognition**, p. 108081, 2021. ISSN 0031-3203. Disponível em: <https://www.sciencedirect.com/science/article/pii/S0031320321002685>. Nenhuma citação no texto.

VIEIRA, P. d. A.; MATHEW, M. J.; NETO, P. d. A. d. S.; SILVA, R. R. V. e. The automated generation of medical reports from polydactyly x-ray images using cnns and transformers. **Applied Sciences**, v. 14, n. 15, 2024. ISSN 2076-3417. Disponível em: <https://www.mdpi.com/2076-3417/14/15/6566>. Nenhuma citação no texto.

VIEIRA, P. de A.; VOGADO, L.; LOPES, L.; RICARDO, R.; NETO, P. S.; MATHEW, M. J.; MAGALHãES, D.; SILVA, R. Deep learning approach for disease detection in lumbosacral spine radiographs using convnet. **Computer Methods in Biomechanics and Biomedical Engineering: Imaging & Visualization**, Taylor & Francis, v. 11, n. 6, p. 2560–2575, 2023. ISSN 2168-1163. Disponível em: <https://doi.org/10.1080/21681163.2023.2245922>. Nenhuma citação no texto.

VIEIRA, P. de A.; VOGADO, L.; LOPES, L.; RICARDO, R.; NETO, P. S.; MATHEW, M. J.; MAGALHãES, D.; SILVA, R. Deep learning approach for disease detection in lumbosacral spine radiographs using convnet. **Computer Methods in Biomechanics and Biomedical Engineering: Imaging & Visualization**, Taylor & Francis, v. 11, n. 6, p. 2560–2575, 2023. Disponível em: <https://doi.org/10.1080/21681163.2023.2245922>. Nenhuma citação no texto.

VINICIUS, A. **Redes Neurais Artificias**. [S.l.], 2017. Disponível em: <https://medium.com/@avinicius.adorno/redes-neurais-artificiais-418a34ea1a39>. Nenhuma citação no texto.

VOGADO, L.; ARAúJO, F.; NETO, P. S.; ALMEIDA, J.; TAVARES, J. M. R.; VERAS, R. A ensemble methodology for automatic classification of chest x-rays using deep learning.

Computers in Biology and Medicine, v. 145, p. 105442, 2022. ISSN 0010-4825. Disponível em: <https://www.sciencedirect.com/science/article/pii/S0010482522002347>. Nenhuma citação no texto.

WANG, X.; PENG, Y.; LU, L.; LU, Z.; BAGHERI, M.; SUMMERS, R. M. Chestx-ray8: Hospital-scale chest x-ray database and benchmarks on weakly-supervised classification and localization of common thorax diseases. **CoRR**, abs/1705.02315, 2017. Nenhuma citação no texto.

WANG, X.; PENG, Y.; LU, L.; LU, Z.; SUMMERS, R. M. Tienet: Text-image embedding network for common thorax disease classification and reporting in chest x-rays. In: **2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition**. [S.l.: s.n.], 2018. p. 9049–9058. Nenhuma citação no texto.

WENG, C.-H.; WANG, C.-L.; HUANG, Y.-J.; YEH, Y.-C.; FU, C.-J.; YEH, C.-Y.; TSAI, T.-T. Artificial intelligence for automatic measurement of sagittal vertical axis using resunet framework. **Journal of Clinical Medicine**, v. 8, p. 1826, 11 2019. Nenhuma citação no texto.

WERNICK, M. N.; YANG, Y.; BRANKOV, J. G.; YOURGANOV, G.; STROTHER, S. C. Machine learning in medical imaging. **IEEE signal processing magazine**, IEEE, v. 27, n. 4, p. 25–38, 2010. Nenhuma citação no texto.

WINESKI, L. **Snell's Clinical Anatomy by Regions**. Wolters Kluwer Health, 2024. ISBN 9781975194123. Disponível em: <https://books.google.fr/books?id=wdbrEAAAQBAJ>. Nenhuma citação no texto.

WOLF, T.; DEBUT, L.; SANH, V.; CHAUMOND, J.; DELANGUE, C.; MOI, A.; CISTAC, P.; RAULT, T.; LOUF, R.; FUNTOWICZ, M.; DAVISON, J.; SHLEIFER, S.; PLATEN, P. von; MA, C.; JERNITE, Y.; PLU, J.; XU, C.; SCAO, T. L.; GUGGER, S.; DRAME, M.; LHOEST, Q.; RUSH, A. Transformers: State-of-the-art natural language processing. In: LIU, Q.; SCHLANGEN, D. (Ed.). **Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: System Demonstrations**. Online: Association for Computational Linguistics, 2020. p. 38–45. Disponível em: <https://aclanthology.org/2020.emnlp-demos.6>. Nenhuma citação no texto.

XIAO, Y.; TIAN, Z.; YU, J.; ZHANG, Y.; LIU, S.; DU, S.; LAN, X. A review of object detection based on deep learning. **Multimedia Tools and Applications**, Springer, v. 79, p. 23729–23791, 2020. Nenhuma citação no texto.

XUE, Y.; TAN, Y.; TAN, L.; QIN, J.; XIANG, X. Generating radiology reports via auxiliary signal guidance and a memory-driven network. **Expert Systems with Applications**, v. 237, p. 121260, 2024. ISSN 0957-4174. Disponível em: <https://www.sciencedirect.com/science/article/pii/S0957417423017621>. Nenhuma citação no texto.

YALING, P.; QIAORAN, C.; TONGTONG, C.; HANQI, W.; XIAOLEI, Z.; ZHIHUI, F.; YONG, L. Evaluation of a computer-aided method for measuring the cobb angle on chest x-rays. **European Spine Journal**, v. 28, 2019. ISSN 1432-0932. Disponível em: <https://doi.org/10.1007/s00586-019-06115-w>. Nenhuma citação no texto.

YENDURI, G.; RAMALINGAM, M.; SELVI, G. C.; SUPRIYA, Y.; SRIVASTAVA, G.; MADDIKUNTA, P. K. R.; RAJ, G. D.; JHAVERI, R. H.; PRABADEVI, B.; WANG, W.; VASILAKOS, A. V.; GADEKALLU, T. R. Gpt (generative pre-trained transformer)—a comprehensive review on enabling technologies, potential applications, emerging challenges, and future directions. **IEEE Access**, v. 12, p. 54608–54649, 2024. Nenhuma citação no texto.

YI, X.; WALIA, E.; BABYN, P. Generative adversarial network in medical imaging: A review. **Medical Image Analysis**, v. 58, p. 101552, 2019. ISSN 1361-8415. Disponível em: <https://www.sciencedirect.com/science/article/pii/S1361841518308430>. Nenhuma citação no texto.

YOSINSKI, J.; CLUNE, J.; BENGIO, Y.; LIPSON, H. **How transferable are features in deep neural networks?** 2014. Nenhuma citação no texto.

ZECH, J. R.; BADGELEY, M. A.; LIU, M.; COSTA, A. B.; TITANO, J. J.; OERMANN, E. K. Variable generalization performance of a deep learning model to detect pneumonia in chest radiographs: A cross-sectional study. **PLOS Medicine**, Public Library of Science, v. 15, n. 11, p. 1–17, 11 2018. Disponível em: <https://doi.org/10.1371/journal.pmed.1002683>. Nenhuma citação no texto.

Zeng, Y.; Liu, X.; Xiao, N.; Li, Y.; Jiang, Y.; Feng, J.; Guo, S. Automatic diagnosis based on spatial information fusion feature for intracranial aneurysm. **IEEE Transactions on Medical Imaging**, v. 39, n. 5, p. 1448–1458, 2020. Nenhuma citação no texto.

ZHANG, J.; XIA, L.; LIU, J.; NIU, X.; TANG, J.; XIA, J.; LIU, Y.; ZHANG, W.; LIANG, Z.; ZHANG, X.; TANG, G.; ZHANG, L. Exploring deep learning radiomics for classifying osteoporotic vertebral fractures in x-ray images. **Frontiers in Endocrinology**, v. 15, 2024. ISSN 1664-2392. Disponível em: <https://www.frontiersin.org/journals/endocrinology/articles/10.3389/fendo.2024.1370838>. Nenhuma citação no texto.

ZHANG, W.; ITOH, K.; TANIDA, J.; ICHIOKA, Y. Parallel distributed processing model with local space-invariant interconnections and its optical architecture. **Appl. Opt.**, OSA, v. 29, n. 32, p. 4790–4797, Nov 1990. Disponível em: <http://ao.osa.org/abstract.cfm?URI=ao-29-32-4790>. Nenhuma citação no texto.

ZHAO, G.; ZHAO, Z.; GONG, W.; LI, F. Radiology report generation with medical knowledge and multilevel image-report alignment: A new method and its verification. **Artificial Intelligence in Medicine**, v. 146, p. 102714, 2023. ISSN 0933-3657. Disponível em: <https://www.sciencedirect.com/science/article/pii/S0933365723002282>. Nenhuma citação no texto.

ZHENG, X.; CHALASANI, T.; GHOSAL, K.; LUTZ, S.; SMOLIC, A. Stada: Style transfer as data augmentation. **Proceedings of the 14th International Joint Conference on Computer Vision, Imaging and Computer Graphics Theory and Applications**, SCITEPRESS - Science and Technology Publications, 2019. Disponível em: <http://dx.doi.org/10.5220/0007353401070114>. Nenhuma citação no texto.

ZHONG, Z.; ZHENG, L.; KANG, G.; LI, S.; YANG, Y. **Random Erasing Data Augmentation**. 2017. Nenhuma citação no texto.

ÇALLı, E.; SOGANCIOGLU, E.; van Ginneken, B.; van Leeuwen, K. G.; MURPHY, K. Deep learning for chest x-ray analysis: A survey. **Medical Image Analysis**, p. 102125, 2021. ISSN 1361-8415. Nenhuma citação no texto.