PUCRS

ESCOLA POLITÉCNICA
PROGRAMA DE PÓS-GRADUAÇÃO EM CIÊNCIA DA COMPUTAÇÃO
MESTRADO EM CIÊNCIA DA COMPUTAÇÃO

DAVIDE CLODE DA SILVA

GERAÇÃO DE IMAGENS DE RAIOS-X DO TÓRAX UTILIZANDO *LATENT DIFFUSION MODELS*

Porto Alegre
2024

PÓS-GRADUAÇÃO - *STRICTO SENSU*

Pontifícia Universidade Católica
do Rio Grande do Sul

PONTIFÍCIA UNIVERSIDADE CATÓLICA DO RIO GRANDE DO SUL
ESCOLA POLITÉCNICA
PROGRAMA DE PÓS-GRADUAÇÃO EM CIÊNCIA DA COMPUTAÇÃO

GERAÇÃO DE IMAGENS DE RAIOS-X DO TÓRAX UTILIZANDO *LATENT DIFFUSION MODELS*

DAVIDE CLODE DA SILVA

Dissertação apresentada como requisito parcial à obtenção do grau de Mestre em Ciência da Computação na Pontifícia Universidade Católica do Rio Grande do Sul.

Orientadora: Prof[a]. Dra. Soraia Raupp Musse

Porto Alegre
2024

Ficha Catalográfica

C643g Clode da Silva, Davide

Geração de Imagens de Raios-x do Tórax Utilizando Latent Diffusion Models / Davide Clode da Silva. – 2024.
105.
Dissertação (Mestrado) – Programa de Pós-Graduação em Ciência da Computação, PUCRS.

Orientadora: Profa. Dra. Soraia Raupp Musse.

1. Foundation Models. 2. Generative Models. 3. Stable Diffusion Model. 4. Latent Diffusion Models. 5. Synthetic Images. I. Musse, Soraia Raupp. II. Título.

Elaborada pelo Sistema de Geração Automática de Ficha Catalográfica da PUCRS
com os dados fornecidos pelo(a) autor(a).
Bibliotecária responsável: Clarissa Jesinska Selbach CRB-10/2051

DAVIDE CLODE DA SILVA

GERAÇÃO DE IMAGENS DE RAIOS-X DO TÓRAX UTILIZANDO *LATENT DIFFUSION MODELS*

Dissertação apresentada como requisito parcial para obtenção do grau de Mestre em Ciência da Computação do Programa de Pós-Graduação em Ciência da Computação, Escola Politécnica da Pontifícia Universidade Católica do Rio Grande do Sul.

Aprovado(a) em 30 de Abril de 2024.

BANCA EXAMINADORA:

Prof. Dr. Thiago Lopes Trugillo da Silveira (INF/UFRGS)

Prof. Dr. Márcio Sarroglia Pinho (PPGCC/PUCRS)

Prof[a]. Dra. Soraia Raupp Musse (PPGCC/PUCRS - Orientadora)

DEDICATÓRIA

Este trabalho é dedicado aos meus pais, Maria e Paulo. À minha mãe, Maria, que mesmo sem ter tido a oportunidade de cursar o ensino superior, nunca poupou esforços para me apoiar em meus estudos e na busca pelos meus objetivos. A determinação e coragem dela diante das adversidades da vida são uma inspiração para mim e me ensinaram a jamais desistir dos meus sonhos. Também dedico este trabalho à minha namorada, Melyssa, que foi compreensiva nos momentos em que precisei me ausentar e me apoiou com palavras de incentivo e carinho durante toda essa jornada.

“*You may encounter many defeats, but you must not be defeated. In fact, it may be necessary to encounter the defeats, so you can know who you are, what you can rise from, how you can still come out of it*.”
(Maya Angelou)

AGRADECIMENTOS

Gostaria de expressar minha profunda gratidão à minha família. Apesar da distância, o incentivo deles foi fundamental para que eu continuasse nessa jornada.

Agradeço ao Brasil, especificamente à Coordenação de Aperfeiçoamento de Pessoal de Nível Superior - CAPES, pelo aporte financeiro que me foi concedido. Este apoio foi fundamental para a realização deste trabalho.

Minha gratidão se estende também à Pontifícia Universidade Católica do Rio Grande do Sul - PUCRS. A jornada acadêmica de alto nível que vivenciei nesta instituição foi inestimável para o meu crescimento pessoal e profissional.

Quero agradecer ao professor Dr. Afonso Henrique Correa de Sales e ao Ministério da Ciência, Tecnologia e Inovações, que, por meio dos recursos da lei n.º 8.248, de 23 de outubro de 1991, no âmbito do PPI-SOFTEX, coordenado pela SOFTEX e publicado Residência em TIC 02 - Aditivo, DOU 01245.012095/2020-56, tornou possível a realização parcial deste trabalho.

Não posso deixar de expressar minha sincera gratidão à minha orientadora, a professora Dra. Soraia Raupp Musse. Sua orientação e leitura minuciosa deste trabalho foram vitais para a sua realização. Além disso, seus feedbacks e sua generosidade como interlocutora, contribuíram significativamente para o meu desenvolvimento acadêmico e para o exito deste trabalho.

Agradeço também ao doutorando Gabriel Fonseca Silva, pela leitura minuciosa e feedbacks ao longo do desenvolvimento deste trabalho. As contribuições das médicas Marina Musse Bernardes e Nathália Giacomini Ceretta foram fundamentais para o desenvolvimento dos testes realizados neste trabalho.

Expresso também minha gratidão ao professor Dr. Rafael Heitor Bordini, cujas ideias contribuíram significativamente para o desenvolvimento deste trabalho.

Por fim, gostaria de agradecer aos professores Dr. Márcio Sarroglia Pinho e Dr. Thiago L. T. da Silveira, pelo pronto aceite em participarem desta banca.

GERAÇÃO DE IMAGENS DE RAIOS-X DO TÓRAX UTILIZANDO *LATENT DIFFUSION MODELS*

RESUMO

A pesquisa em imagens médicas é uma tarefa desafiadora, devido, mas não se limitando, à escassez de imagens médicas. Usualmente, dados reais são usados na realização de simulações. Contudo, diversos fatores limitam a capacidade de utilizar dados de exames médicos. Entre eles, destacam-se a necessidade de proteger informações sensíveis e confidenciais dos pacientes, o custo e a acessibilidade dos equipamentos médicos, além de questões éticas e legais relacionadas ao compartilhamento de dados. Para mitigar esses desafios, a geração de dados sintéticos oferece uma alternativa promissora, permitindo complementar conjuntos de dados de treinamento e realizar pesquisas de imagens médicas em larga escala. Recentemente, *Diffusion Models* ganharam destaque na comunidade de visão computacional ao produzir imagens sintéticas fotorrealistas. Neste contexto, este trabalho propõe explorar o uso de um tipo específico de *Diffusion Models*, os *Latent Diffusion Models* (LDM), na geração de imagens médicas sintéticas a partir de imagens torácicas de alta resolução. Para isso, realizamos um ajuste fino (*fine-tuning*), no *Latent Diffusion Foundation Model*, utilizando imagens da *National Library of Medicine*. Como contribuição, geramos um conjunto de imagens de raios-x do tórax saudáveis de alta resolução e excelente qualidade. Além disso, disponibilizamos essas imagens geradas juntamente com os modelos treinados para serem utilizadas em pesquisas acadêmicas ou no treinamento de outros modelos. Sendo assim, a nossa abordagem oferece potencial para superar a barreira dos dados no domínio médico.

Palavras-Chave: *Foundation Models*, *Generative Models*, *Stable Diffusion Model*, *Latent Diffusion Models*, *Generative Adversarial Networks*, Imagens Sintéticas.

CHEST X-RAYS GENERATION USING LATENT DIFFUSION MODELS

ABSTRACT

Medical imaging research is a challenging task due to, but not limited to, the scarcity of medical images. Usually, real data is used to carry out simulations. However, several factors limit the ability to use medical exam data. These include the need to protect sensitive and confidential patient information, the cost and accessibility of medical equipment, as well as ethical and legal issues related to data sharing. To mitigate these challenges, synthetic data generation offers a promising alternative, allowing to complement training datasets and conduct large-scale medical imaging research. Recently, *Diffusion Models* have gained prominence in the computer vision community by producing photorealistic synthetic images. In this context, this work proposes to explore the use of a specific type of *Diffusion Models*, the *Latent Diffusion Models* (LDM), in generating synthetic medical images from high-resolution thoracic images. To do this, we performed a fine-tuning on the *Latent Diffusion Foundation Model*, using images from the *National Library of Medicine*. As a contribution, we generated a set of high-resolution, excellent-quality healthy chest x-ray images. Additionally, we provide these generated images alongside the trained models for use in academic research or for training other models. As such, our approach offers the potential to overcome the data barrier in the medical domain.

Keywords: Foundation Models, Generative Models, Stable Diffusion Model, Latent Diffusion Models, Generative Adversarial Networks, Synthetic Images.

LISTA DE FIGURAS

LISTA DE TABELAS

LISTA DE SIGLAS

SDM – *Stable Diffusion Model*

LDM – *Latent Diffusion Model*

VAE_S – *Variational Autoencoders*

ARM*s* – *Autoregressive Models*

GAN*s* – *Generative Adversarial Networks*

DPM*s* – *Diffusion Probabilistic Models*

LDFM – *Latent Diffusion Foundation Model*

LORA – *Low-Rank Adaptation*

FM*s* – *Foundation Models*

CF – *Catastrophic forgetting*

MRI – *Magnetic Resonance Imaging*

CT – *Computed Tomography*

OCT – *Optical Coherence Tomography*

PET – *Positron Emission Tomography*

IR – *Information Retrieval*

QA – *Question Answering*

KB – *Knowledge Base*

CLIP – *Contrastive Language-Image Pre-training*

ELMO – *Embeddings from Language Models*

BERT – *Bidirectional Encoder Representations from Transformers*

KBC – *Knowledge Base Construction*

SDMS – *Stable Diffusion Models*

SUMÁRIO

1. INTRODUÇÃO

Machine Learning (ML) tem desempenhado papel fundamental ao longo de anos na área de saúde. No entanto, o seu impacto nesta área tem sido incontestavelmente limitado comparado a outros domínios de aplicação. Por exemplo, na prevenção e tratamento de doenças, ML pode auxiliar na análise de grandes conjuntos de dados para identificar tendências e fazer previsões sobre os resultados das doenças, como modelar a progressão e o tratamento de condições cancerígenas [60]. No entanto, a adoção de técnicas de ML nos cuidados de saúde tem sido lenta devido a fatores como a escassez de dados de pacientes, preocupações com a privacidade dos dados, requisitos regulamentares e a natureza crítica das decisões em matéria de cuidados de saúde [33, 100]. Por outro lado, em domínios como fabricação e cadeia de suprimentos, o uso de técnicas de ML tem trazido benefícios significativos possibilitando a geração de inteligência acionável a partir do processamento de dados coletados, o que aumenta a eficiência da fabricação sem necessitar de alterações significativas nos recursos. Além disso, a capacidade das técnicas de ML de fornecer *insights* preditivos permite discernir padrões de fabricação complexos e oferece um caminho para um sistema inteligente de suporte à decisão em uma variedade de tarefas de fabricação, como inspeção inteligente e contínua, manutenção preditiva, melhoria de qualidade, otimização de processos, gerenciamento da cadeia de suprimentos e agendamento de tarefas [84].

A garantia de proteção de dados dos pacientes devido às informações sensíveis e confidenciais que neles constam, e que exigem proteção contra acessos não autorizados, são um dos fatores que dificultam o acesso aos dados médicos dos pacientes. Além disso, a falta de padronização dos registros e o esforço de coletar esses dados médicos também representam desafios da área. A possibilidade de geração de dados médicos sintéticos, realistas e de alta qualidade, pode ser uma alternativa plausível para mitigar alguns desses desafios. De maneira geral, a indústria prevê um aumento significativo na disponibilidade de dados sintéticos nos próximos anos, podendo ser até duas vezes mais abundantes do que os dados reais atualmente disponíveis [1]. Alguns exemplos são modelos generativos para a síntese de imagens fotorrealísticas a partir de uma determinada descrição em linguagem natural [86] e também para melhorar o desempenho de modelos baseados nas GANs, em imagens de super-resolução [28].

Especificamente no contexto de imagens médicas, o uso de modelos generativos pode constituir uma solução viável. Na comunidade acadêmica, tem havido algumas explorações iniciais tentando fazer *fine-tuning* de *Foundation Models* (re-treinamento de modelos pré-treinados) usando pequenas quantidades de dados, para poderem ser generalizados de forma eficiente para aplicações específicas. Um exemplo notável é o projeto

[1] https://www.eckerson.com/articles/synthetic-data-for-ai-definition-risks-and-strategies

denominado MedCLIP [104]. Neste projeto, um modelo CLIP foi ajustado no conjunto de dados denominado ROCO, composto por imagens radiológicas e suas legendas correspondentes [104, 83]. Segundo os pesquisadores responsáveis, o modelo ajustado, MedCLIP, é capaz de identificar características de nível superior, como a modalidade de uma imagem. Por exemplo, os autores afirmam que o modelo é capaz de determinar se uma determinada imagem radiológica é uma PET ou uma ultrassonografia. No entanto, o MedCLIP atualmente apresenta limitações na identificação de características mais específicas. Ele ainda não é capaz de distinguir com precisão uma tomografia cerebral de uma tomografia pulmonar [104]. Ainda na literatura, há trabalhos que utilizam a abordagem de modelos generativos para diferentes finalidades, por exemplo, para a síntese de dados médicos realísticos [73, 82] e para síntese de dados visando o aumento de conjunto de dados de treinamento de modelos de aprendizado profundo [28].

1.1 OBJETIVOS

O principal objetivo deste trabalho é utilizar um método capaz de gerar imagens médicas sintéticas de alta resolução e precisas. Propomos a adoção da abordagem da Inteligência Artificial Generativa, onde um modelo pré-treinado (*Foundation Model*) é utilizado como base para realizar *fine-tuning*.

Para alcançar o nosso objetivo principal, estabelecemos os seguintes objetivos específicos:

- Realizar *fine-tuning* do *Latent Diffusion Foundation Model* utilizando imagens e relatórios clínicos de raios-x do tórax para melhorar a precisão e a eficácia do modelo.
- Empregar a ferramenta Kohya-ss [2] para auxiliar na configuração de parâmetros treináveis, visando otimizar o desempenho do modelo.
- Aplicar *Latent Diffusion Model* (LDM) para gerar imagens médicas sintéticos de raios-x do tórax de alta resolução.
- Gerar imagens de raios-x do tórax usando a interface AUTOMATIC1111 [3] como ferramenta auxiliar.
- Analisar as imagens de raios-x do tórax sintéticos geradas e avaliar suas qualidades com a participação de especialistas na área, visando garantir a precisão e a utilidade dos dados gerados.

[2] https://github.com/bmaltais/kohya_ss
[3] https://github.com/AUTOMATIC1111/stable-diffusion-webui

1.2 Principais Contribuições

A principal contribuição deste trabalho é a geração de um conjunto de imagens de raios-x do tórax saudáveis de alta resolução e avaliadas por médicos como sendo de excelente qualidade. Além disso, exploramos o uso do método proposto, o *Stable Diffusion Model*, realizando o ajuste fino (*fine-tuning*) utilizando imagens de raios-x do tórax reais. Como resultado, disponibilizamos modelos treinados com diferentes tamanhos de conjuntos de dados, juntamente com as imagens de raios-x do tórax geradas, para uso em pesquisas acadêmicas e no treinamento de outros modelos [4].

1.3 ORGANIZAÇÃO DO TRABALHO

Este trabalho está organizado da seguinte forma: O Capítulo 2 apresenta os trabalhos relacionados que desempenharam um papel importante na compreensão de diferentes abordagens utilizadas na geração de dados médicos sintéticos. O Capítulo 3 discorre sobre o método proposto. O Capítulo 4 traz os resultados e as discussões gerais sobre os resultados alcançados. Por fim, o Capítulo 5 apresenta as considerações finais, as limitações e possíveis contribuições futuras.

[4] https://github.com/DavideDaSilva/MESTRADO-PUCRS.git

2. TRABALHOS RELACIONADOS

A geração de imagens usando modelos de Inteligência Artificial (IA) é uma área de estudo em constante evolução. O desafio consiste em utilizar essas tecnologias para produzir imagens de alta qualidade, precisas e em grande escala. Diversas metodologias têm sido exploradas na literatura, como o uso de *Foundation Model* LAFITE [117] e o Re-Imagen [20], entre outras.

O *Foundation Model* LAFITE tem recebido atenção considerável na geração de imagens, apresentando resultados promissores. Zhou et al. [117] propuseram o LAFITE (*LAnguage-Free traIning for Text-to-image gEneration*) para resolver um dos maiores desafios no treinamento de modelos de geração de imagens (modelos *text-to-image*), que é a necessidade de um grande número de pares imagem-texto. De acordo com os autores, O modelo proposto pode ser treinado sem quaisquer dados de texto e tem demonstrado resultados promissores. Os autores ainda afirmam que o método pode ser aplicado no ajuste fino (*fine-tuning*) de modelos pré-treinados, economizando tempo e custos de treinamento.

Por outro lado, o *Foundation Model Re-Imagen*, também conhecido como *Retrieval-Augmented Text-to-Image Generator*, é um modelo generativo que utiliza informações recuperadas para produzir imagens precisas. Esse modelo se destaca por sua capacidade de gerar imagens fiéis, mesmo para entidades raras ou nunca antes vistas [20].

Os *Foundation Models* são capazes de aprender a partir de uma abundância de dados diversos, o que lhes permite generalizar e abstrair dados para novas tarefas, com quantidades de dados relativamente pequenas e específicas da tarefa. Isto pode ser útil em contextos de análise de imagens médicas, onde a obtenção de grandes conjuntos de dados rotulados pode constituir uma tarefa desafiadora devido a questões de privacidade e ao esforço necessário para a anotação especializada [6].

2.1 *Foundation Models*

Foundation Models (FMs) ou Modelo de Base, são modelos de *Machine Learning* treinados em um amplo espectro de dados generalizados e não rotulados, usando geralmente a técnica de autossupervisão, e podem ser adaptados a amplas tarefas [10]. Alguns exemplos de *Foundation Models* incluem ELMo [80], BERT [54], GPT-3 [11], CLIP [83], ResNet [39], DALL-E [85] e *Stable Diffusion* [89]. Esses modelos obtiveram ganhos substanciais em diversas tarefas desafiadoras, como resposta a perguntas (do inglês *Question Answering (QA)*) [11], construção de base de conhecimento (do inglês *Knowledge Base*

Construction (KBC)) [81] e recuperação de informações (do inglês *Information Retrieval* (IR)) [36].

Foundation Models têm o potencial de transformar vários sectores, ampliando o papel que a IA desempenha na sociedade. Dentre as inúmeras aplicações nas quais os *Foundation Models* podem ser empregados, destacam-se três áreas – saúde, direito e educação – que são áreas fundamentais para a função social [10]. A aplicação de *Foundation Models* nas áreas elencadas traz consigo, além de oportunidades que eles representam, os desafios (interpretabilidade) e as preocupações (privacidade, segurança, etc.) [10]. No entanto, este trabalho não irá explorar todas elas. Em vez disso, concentrar-se-á na aplicação de *Foundation Models* na área da saúde, a qual é o foco principal deste trabalho.

2.1.1 Oportunidades na Área da Saúde

Os cuidados de saúde, como a investigação biomédica, demandam despesas significativas, tempo e conhecimento médico abrangente. No trabalho intitulado *"Allocation of Physician Time in Ambulatory Practice: A Time and Motion Study in 4 Specialties*", os autores concluíram que para cada hora que os médicos fornecem atendimento clínico direto aos pacientes, quase duas horas adicionais são gastas no registro eletrônico de saúde (*Electronic Health Record- EHR*) e trabalho administrativo durante *Day Clinic* (refere-se a uma modalidade de atendimento médico na qual o paciente é admitido por apenas um dia para realizar procedimentos clínicos ou cirúrgicos, com risco reduzido de infecção). Fora do horário de expediente, os médicos passam mais uma a duas horas do tempo pessoal todas as noites fazendo trabalhos adicionais de computador e outros trabalhos administrativos [94].

Os *Foundation Models*, no contexto da saúde, podem desempenhar um papel fundamental, uma vez que são considerados repositórios centrais de conhecimento médico treinado em diversas fontes ou modalidades de dados médicos, desde resultados de ensaios clínicos até registros de saúde de pacientes e dados genômicos, consultados ou atualizados interativamente por profissionais médicos e pelo público [61, 29].

Além disso, os *Foundation Models* têm capacidade adaptativa, uma vez que podem ser adaptados de forma eficiente a várias tarefas individuais em saúde e biomedicina. Entre as capacidades adaptativas podem ser destacados o aplicativo de automatização de resposta a perguntas usado por pacientes. Citam-se como exemplos os aplicativos de registro em diário que permitem aos usuários registrar dados sobre sua dieta, exercícios, níveis de glicose no sangue, e outros comportamentos e medidas relacionadas com a saúde; terminais de pacientes para telemonitoramento de condições como hipertensão e insuficiência cardíaca crônica; aplicações que recebem dados de pedômetros, monitores de pressão arterial e outros dispositivos; jogos que ensinam habilidades relacionadas

à saúde, e assim por diante [23, 58] e o sistema automatizado de correspondência de ensaios clínicos que utiliza processamento de linguagem natural para extrair características de pacientes e ensaios de fontes não estruturadas e aprendizado de máquina para combinar pacientes com testes clínicos [7].

Com o aumento da expectativa de vida e o crescimento populacional, a demanda por serviços de saúde tende a crescer a um ritmo sem precedentes [115]. Além disso, à medida que essa demanda aumenta, a sociedade tende a enfrentar uma falta de profissionais de saúde [57]. Esta ineficiência, juntamente com o contínuo crescimento populacional e a falta de profissionais de saúde, exige o desenvolvimento de interfaces precisas, como sistemas automatizados, para auxiliar no diagnóstico e tratamento de doenças, na síntese de registros de pacientes e na resposta a perguntas dos pacientes [24, 75, 103] *apud* [10].

Os *Foundation Models* têm o potencial para aprimorar a assistência ao paciente através dos profissionais de saúde. Eles podem, a título de exemplo, aprimorar a compreensão do estado clínico do paciente por meio da identificação, monitoramento e avaliação, utilizando dados provenientes de diversas fontes como prontuários eletrônicos, sensores e dispositivos vestíveis. Além disso, os *Foundation Models* podem contribuir para a diminuição da carga administrativa do médico, auxiliando na síntese e organização da informação clínica [113].

As descobertas biomédicas, como o desenvolvimento de medicamentos, que envolve processos complexos que começam desde a pesquisa básica de medicamentos, identificação de alvos proteicos e a descoberta de moléculas potentes, até o desenvolvimento clínico (ensaios clínicos), demandam recursos humanos, tempo experimental e custos financeiros significativos [107]. Facilitar e acelerar essas descobertas biomédicas usando dados existentes e descobertas publicadas são tarefas desafiadoras na biomedicina [112]. Os *Foundation Models* podem ser úteis para essas descobertas biomédicas nos seguintes aspectos:

- Primeiro, através de sua capacidade generativa (como a geração de textos utilizando GPT-3), os *Foundation Models* podem auxiliar nas tarefas generativas em pesquisa biomédica (ensaios clínicos) e projetar moléculas que funcionem (descoberta de medicamentos) com base nos dados existentes [38, 51].

- Segundo, devido ao seu potencial em integrar diversas modalidades de dados médicos, os *Foundation Models* podem auxiliar na investigação de conceitos biomédicos em múltiplas escalas (usando dados ao nível de molécula, paciente e população) e múltiplas fontes de conhecimento (usando imagens e textos). Isto facilita descobertas biomédicas que são difíceis de obter se forem utilizados dados de modalidade única [1, 59, 62, 87, 90, 102, 109].

2.1.2 Desafios dos *Foundations Models*

Apesar da existência de amplas oportunidades para a geração de dados sintéticos por meio dos *Foundation Models*, estes também trazem consigo desafios significativos. Estes desafios são particularmente notáveis nas áreas de saúde e biomédicas, o que estimula um interesse considerável na condução de pesquisas que abordam uma variedade de temas. Entre estes desafios, pode-se citar:

Multimodalidade: Os dados médicos são altamente multimodais, podendo envolver diferentes tipos de dados (texto, imagem, vídeo, etc.), escalas (molécula, gene, célula, etc.) [59, 90] e estilos (linguagem profissional e leiga) [63, 65] *apud* [10]. Os atuais modelos autossupervisionados são desenvolvidos para cada modalidade. A título de exemplo, podemos citar a imagem [13], texto [64], gene [47], proteína [50], e não aprendem conjuntamente em diversas modalidades. Desta forma, para aprender as informações intermodais e crossmodais torna-se necessário investigar estratégias de fusão ao nível de recurso e no nível semântico no treinamento de *Foundation Models* [10].

Explicabilidade: Trata-se de fornecer evidências e etapas lógicas para a tomada de decisões – ela é crucial na área da saúde e na biomedicina [41]. Por exemplo, em diagnósticos e ensaios clínicos, os sintomas do paciente e a relevância temporal devem ser explicados como evidência. Isso ajuda na resolução de possíveis divergências entre o sistema e especialistas humanos. A explicabilidade também é necessária para o consentimento informado na área de saúde [3]. No entanto, os objetivos de treinamento dos atuais *Foundation Models* não incluem explicabilidade, exigindo, desta forma, pesquisas futuras nesta direção [67]. A incorporação de gráficos de conhecimento pode ser um passo plausível para melhorar ainda mais a explicabilidade do modelo [88, 110, 49] *apud* [10].

Regulamentações legais e éticas: As aplicações de saúde exigem observância de regulamentações legais e éticas com garantias, como segurança do paciente, privacidade e justiça. Por exemplo, no que diz respeito à segurança, as previsões feitas pelos *Foundation Models* devem ser factualmente precisas com o conhecimento médico estabelecido e devem quantificar a incerteza ou optar por submeter-se a um especialista quando incerto [14, 74] *apud* [10]. No quesito privacidade, o uso de registros de saúde dos pacientes deve observar as leis de proteção de dados [16] *apud* [10]. Quanto à justiça, os pesquisadores devem estar cientes das armadilhas comuns para evitar o risco de exacerbar as desigualdades sociais existentes [19, 105, 18]. Além disso, deve ser garantido que os dados de formação e avaliação dos *Foundation Models* sejam suficientemente representativos dos diferentes sexos,

raças, etnias e origens socioeconômicas [52, 71] *apud* [10]. Quando os dados representativos são escassos, é necessário pesquisas para desenviesar e regularizar modelos para garantir justiça [116] *apud* [10]. Por fim, os desenvolvedores de *Foundation Models* também devem consultar pesquisadores de ética e direito e observar os regulamentos nas circunstâncias específicas (por exemplo, país, região) onde são implantados [10].

Extrapolação: O processo de descoberta biomédica envolve extrapolação. Por exemplo, os *Foundation Models* devem ser capazes de se adaptarem rapidamente a novas tecnologias experimentais (por exemplo, novos ensaios, novas técnicas de imagem, como microscopia de alta resolução) ou novos ambientes (por exemplo, novas doenças alvo) [8, 46] *apud* [10]. A esse processo de descoberta biomédica dá-se o nome de extrapolação. No entanto, a capacidade de aproveitar conjuntos de dados existentes e extrapolar para novos ambientes é um desafio de aprendizado de máquina na biomedicina [95, 69] *apud* [10].

2.1.3 *Foundation Model Fine-tuning*

Fine-tuning está fortemente associado ao *Transfer Learning* (aprendizagem por transferência) [42]. *Transfer learning* (TL) consiste na aplicação do conhecimento adquirido na resolução de um problema a um problema distinto, mas correlato [79, 97]. Por outro lado, *Fine-tuning* pode ser entendido como uma forma de aplicar, implementar ou utilizar *Transfer Learning*. Trata-se de um processo pelo qual um modelo previamente treinado para uma determinada tarefa, é ajustado a fim de que ele possa executar uma tarefa semelhante, mas diferente [42]. Essa técnica é importante, uma vez que permite utilizar uma rede neural artificial previamente projetada e treinada sem a necessidade de ter que desenvolvê-la do zero.

Os *Foundation Modelos* multimodais são capazes de aprender conceitos visuais e incorporar recursos de imagem e texto em um espaço semântico compartilhado, obtendo assim a capacidade de previsão de *zero-shot*. Ocorre que, quando os dados da tarefa *downstream* pertencem a alguns domínios específicos, como sensoriamento remoto, saúde, etc., que diferem muito dos dados de pré-treinamento, a precisão da previsão *zero-shot* dos *Foundation Models* cai drasticamente [68]. Neste ponto, é necessário fazer *fine-tuning* do modelo com a ajuda de dados da tarefa *downstream*, usando métodos como *linear probing* ou *fine-tuning* global. Porém, esses métodos às vezes requerem um número grande de amostras para um treinamento eficaz, e o número de amostras disponíveis na tarefa *downstream* é limitado às vezes. Para solucionar esse problema, há importantes pesquisas na comunidade acadêmica tentando fazer *fine-tuning* na *Founda-*

tion Model, usando pequenas quantidades de dados para poderem ser generalizados de forma eficiente para aplicações específicas [32, 66, 68].

2.2 Modelos Generativos

Esta seção é dedicada à apresentação de uma variedade de estudos relacionados à geração de imagens sintéticas de raios-x do tórax. Esses estudos empregam diversas abordagens, entre as quais se destacam o uso de Redes Adversárias Generativas (*Generative Adversarial Networks - GANs*) e Modelos de Difusão Latente (*Latent Diffusion Models - LDMs*). Ambas as abordagens têm demonstrado resultados promissores na geração de imagens de raios-x do tórax, com destaque para os LDMs, contribuindo significativamente para o avanço da pesquisa nesta área. A seguir, discutiremos em detalhes essas abordagens e os estudos que as utilizam.

Os métodos descritos aplicam uma ampla variedade de técnicas para alcançar seus objetivos, sendo que a maioria deles recorre a dados do mundo real [96].

Embora os métodos encontrados na literatura apresentem diferentes etapas ou estruturas para a geração de imagens sintéticas de raios-x do tórax, muitos deles seguem uma estrutura que se baseia no trabalho de Rombach et al. [89]. Esta abordagem comum demonstra a influência significativa deste trabalho no campo da geração de imagens médicas sintéticas.

Ao longo de anos, a geração de imagens sintéticas foi principalmente explorada por meio de quatro categorias de modelos generativos: *Variational Autoencoders (VAEs)*, *Autoregressive Models (ARMs)*, *Generative Adversarial Networks (GANs)* e *Diffusion Probabilistic Models (DPMs)* [111]. Todavia, neste estudo, será feito uma breve apresentação desses modelos e da contribuição dos pesquisadores da área, com um foco especial nos *Diffusion Probabilistic Models*.

2.2.1 *Variational Autoencoders (VAEs)*

Variational Autoencoders (VAEs) são modelos de aprendizado autos-supervisionado onde, durante o treinamento, o *output* é uma aproximação do *input*. Normalmente, os *autoencoders* são constituídos por três partes: ***Encoder*** (que produz uma representação compacta do espaço latente dos dados de entrada), o **Espaço Latente** (que retém a semântica dos dados de entrada com dimensionalidade reduzida, mas preserva o máximo de informações) e o ***Decoder*** (que reconstrói os dados compactados vindos do espaço latente).

Em parte, os VAEs têm uma certa vantagem comparado aos modelos GANs, pois são mais estáveis no treinamento, não sofrem com o problema de modo colapso (que é um fenômeno que ocorre quando o Gerador só é capaz de gerar um número limitado de imagens de saída que enganam o Discriminador) e são melhores na estimativa de distribuição de probabilidade. Por outro lado, as Redes Adversárias Generativas (GANs) tendem a superar os modelos VAEs em fidelidade de imagem [93]. Quanto aos modelos autorregressivos (ARMs), estes são eficazes na estimativa de densidade. Contudo, como suas arquiteturas exigem uma alta capacidade computacional e o seu processo de amostragem ser sequencial, os limitam a gerar imagens de baixa resolução. São, portanto, mais lentos que os VAEs [93, 106].

Cetin et al. [12], propuseram a abordagem Attri-VAE para analisar dados de pacientes saudáveis e com infarto do miocárdio. Os autores justificaram a necessidade dessa abordagem, argumentando que são necessários métodos de aprendizagem profunda em que a interpretabilidade seja intrinsecamente considerada como parte do modelo para compreender melhor a relação dos atributos clínicos e baseados em imagens com os resultados da aprendizagem profunda. Os autores ainda argumentam que essa compreensão pode facilitar seu uso no raciocínio por trás das decisões médicas.

Por outro lado, Chatterjee et al. [17], no trabalho intitulado *"Variational Autoencoder Based Imbalanced COVID-19 Detection Using Chest X-Ray Images"*, apresentaram uma nova abordagem para enfrentar o desafio de conjuntos de dados médicos altamente desbalanceados na detecção de COVID-19. Os autores propuseram o método VAEs onde, primeiro, as imagens de radiografia de tórax são convertidas em um espaço latente, aprendendo as características mais importantes usando VAEs. Em segundo lugar, é utilizada outras técnicas de reamostragem de dados para balancear as classes desbalanceadas preexistentes. Por fim, o conjunto de dados modificado é usado para treinar modelos de classificação para classificar imagens de radiografia de tórax em três classes diferentes: COVID-19, Pneumonia e Normal.

2.2.2 *Autoregressive Models (ARMs)*

Os modelos autorregressivos (ARMs) são poderosos na estimativa de densidade. Entretanto, como suas arquiteturas exigem uma alta capacidade computacional e seus processos de amostragem serem sequenciais, os restringem a gerar imagens de baixa resolução. Portanto, isso os torna mais lentos na amostragem em comparação com os VAEs [93, 106].

Apesar das limitações apresentadas, há trabalhos desenvolvidos utilizando a abordagem ARMs. Por exemplo, Han et al. [37], propuseram MedGen3D, uma estrutura generativa baseada em difusão, para sintetizar imagens médicas 3D. Os autores

argumentam que adquirir e anotar (rotular) dados suficientes é importante no desenvolvimento de modelos robustos e precisos baseados em aprendizagem, porém a obtenção de tais dados pode ser um desafio em muitas tarefas de segmentação de imagens médicas. Os autores ainda argumentam que abordagem proposta trata dados médicos como sequências de fatias e emprega um processo autorregressivo para gerar imagens 3D sequencialmente.

Tudosiu et al. [101], propuseram o dimensionamento e otimização de modelos *Vector Quantised-Variational Autoencoder* (VQ-VAE) e *Transformer* para geração de dados de alta resolução. Na abordagem proposta, os autores apresentaram um modelo generativo para a geração de imagens do cérebro humano precisas e de alta resolução. O trabalho foi motivado pelos desafios da escassez de dados e das limitações computacionais que dificultam a geração de imagens volumétricas 3D de alta resolução. Os autores ainda sugerem que o modelo tem o potencial de possibilitar estudos da anatomia e patologia humana em larga escala sem comprometer a privacidade do paciente.

2.2.3 *Generative Adversarial Networks* (GAN*s*)

As Redes Adversárias Generativas (GANs) têm sido, outrora, a abordagem dominante utilizada na geração de imagens. Embora os modelos GANs apresentem resultados promissores para dados com variabilidade limitada, eles apresentam vários problemas. Um deles é o *Catastrophic forgetting (CF)* ou esquecimento catastrófico que é um problema em que o conhecimento de tarefas previamente aprendidas é abruptamente destruído pelo aprendizado da tarefa atual. Quando uma rede GAN sofre de CF, ela exibe comportamentos indesejados, como o **modo colapso** e a **não convergência**. Além disso, os redes GANs também têm dificuldade para capturar a distribuição completa de dados [99].

As redes GANs foram propostas em 2014 por GOODFELLOW et al. [34]. Elas são baseadas em um cenário da teoria de jogos chamado minimax. Neste cenário, um **Discriminador**, denotado por **D**, e um **Gerador**, denotado por **G**, competem entre si. O Gerador gera dados a partir de ruído estocástico (aleatório) e o Discriminador tenta dizer ou distinguir se os dados gerados são reais (vindos de um conjunto de treinamento) ou gerados (vindos da rede geradora) [77, 34]. O Gerador tenta enganar o Discriminador gerando imagens de aparência realística a partir de um vetor latente aleatório, e o Discriminador tenta distinguir a imagem gerada da imagem real [21]

A estrutura de modelagem das GANs, proposta por GOODFELLOW et al. [34], é definida como mostrada na função objetiva da Figura 2.1 abaixo.

Esta função representa a função objetiva do processo de treinamento das GANs. Ela é constituída por duas partes principais: o Gerador (G) e o Discriminador (D). Vamos decompô-la em partes para o melhor entendimento:

$$\min_G \max_D V(D,G) = \mathbb{E}_{\boldsymbol{x}\sim p_{\text{data}}(\boldsymbol{x})}[\log D(\boldsymbol{x})] + \mathbb{E}_{\boldsymbol{z}\sim p_{\boldsymbol{z}}(\boldsymbol{z})}[\log(1 - D(G(\boldsymbol{z})))]$$

Figura 2.1: Função objetiva do processo de treinamento das GANs [34].

A Figura 2.2 representa que o Gerador (G) e o Discriminador (D) possuem objetivos opostos. Enquanto o Gerador tenta minimizar a função, o Discriminador busca maximizá-la. Por outro lado, a Figura 2.3 corresponde à primeira parte da função de valor $V(D, G)$. A expectativa (E) é assumida em todas as instâncias de dados reais (x) extraídas da distribuição de dados verdadeira $p_{\text{data}}(x)$. O termo $\log D(x)$ representa o logaritmo da saída do Discriminador para uma instância de dados real. O objetivo do Discriminador é maximizar esse valor, ou seja, atribuir alta probabilidade a instâncias de dados reais. Já a Figura 2.4 corresponde à segunda parte da função de valor $V(D, G)$. A expectativa (E) é calculada sobre todas as variáveis de ruído z extraídas da distribuição de ruído $p_z(z)$. O termo $\log(1 - D(G(z)))$ representa o logaritmo de um menos a saída do Discriminador para uma instância de dados sintéticos gerada pelo Gerador. O objetivo do Gerador é minimizar esse valor, ou seja, deseja que o Discriminador atribua alta probabilidade às suas instâncias de dados sintéticos (equivalentemente, um valor baixo para $1 - D(G(z))$).

$$\min_G \max_D V(D, G)$$

Figura 2.2: Função objetiva do processo de treinamento das GANs.

$$\mathbb{E}_{\boldsymbol{x}\sim p_{\text{data}}(\boldsymbol{x})}[\log D(\boldsymbol{x})]$$

Figura 2.3: Função objetiva do processo de treinamento das GANs.

$$\mathbb{E}_{\boldsymbol{z}\sim p_{\boldsymbol{z}}(\boldsymbol{z})}[\log(1 - D(G(\boldsymbol{z})))]$$

Figura 2.4: Função objetiva do processo de treinamento das GANs.

Em resumo, o processo adversário entre o Gerador (G) e o Discriminador (D) envolve objetivos opostos. O Discriminador busca maximizar a função total, ou seja, classificar corretamente as instâncias de dados reais e sintéticas. Por outro lado, o Gerador visa minimizar a segunda parte da função, enganando o Discriminador para que este classifique incorretamente suas instâncias de dados sintéticos como verdadeiros. Esse embate leva o Gerador a aprender a gerar dados que são indistinguíveis dos dados reais.

Os autores treinaram a rede GAN em uma variedade de conjuntos de dados, incluindo MNIST, Toronto Face Database (TFD) e CIFAR-10. As redes geradoras utilizaram uma combinação de funções de ativação lineares retificadoras e funções de ativação sigmóides, enquanto a rede discriminadora empregou funções de ativação Maxout . Além disso, a técnica de dropout foi aplicada no treinamento da rede discriminadora. Embora a estrutura teórica permita o uso de dropout e outros ruídos nas camadas intermediárias do gerador, os autores optaram por utilizar o ruído apenas como entrada para a camada mais inferior da rede geradora [34]. A seguir, apresentamos amostras de imagens sintéticas geradas por GANs treinadas no conjunto de dados Toronto Face Database (TFD) (Figura 2.5).

Figura 2.5: Visualização de amostras geradas pelo modelo. As amostras apresentadas no artigo foram escolhidas aleatoriamente pelos autores [34].

Sundaram et al. [96] propuseram o uso de GANs para aumentar artificialmente o conjunto de dados CheXpert de radiografias de tórax, visando lidar com classes desbalanceadas. Nesse contexto, os autores desenvolveram uma estratégia que envolve a criação de imagens sintéticas de raios-x do tórax por meio de Redes Adversárias Generativas (GANs). Essas imagens sintéticas incorporam pelo menos uma das três patologias sub-representadas: lesão pulmonar, pleural ou fratura. Posteriormente, as imagens geradas pela GAN são adicionadas ao conjunto de dados original do CheXpert, contribuindo para a redução do desequilíbrio entre as classes. A Figura 2.6 ilustra esse processo.

A seguir, apresentamos a descrição de cada uma das etapas do processo de aumento de dados apresentado na Figura 2.6:

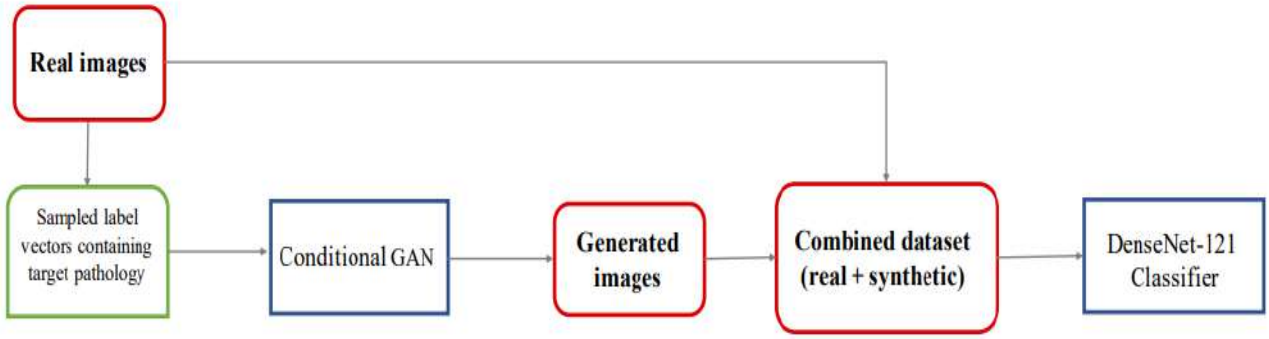


Figura 2.6: Ilustração do método de aumento de dados proposto baseado em GAN. O método combina imagens geradas por GAN com imagens reais de patologias sub-representadas para produzir um conjunto de dados aumentado que pode ser usado para treinamento [96].

- *Real images* (Imagens reais): Essas são imagens de radiografias de tórax reais do conjunto de dados, ou seja, representam dados autênticos.
- *Sampled label vectors containing target pathology*: os autores selecionam vetores de rótulos do conjunto de dados reais que correspondem a patologias específicas, como lesões pulmonares, pleurais ou fraturas. Esses vetores indicam se uma patologia específica está presente em uma imagem de raio-X.
- *Conditional GAN*: Uma GAN é usada para gerar imagens sintéticas. O gerador recebe como entrada um "vetor rótulo" ou vetor de classes (indicando a presença ou ausência de patologias) e produz uma radiografia de tórax sintética. O discriminador avalia o quão bem a imagem gerada corresponde às imagens reais.
- *Generated images*: A rede GAN gera radiografias sintéticas de tórax com patologias específicas. Essas imagens sintéticas são difíceis de distinguir das reais.
- *Combined dataset (real + synthetic)*: As imagens geradas são combinadas com as imagens reais originais. Esse conjunto de dados aumentado contém exemplos reais e sintéticos, visando reduzir o desequilíbrio de classes, introduzindo mais casos de patologias sub-representadas.
- *DenseNet-121 Classifier*: (a) O conjunto de dados aumentado é usado para treinar uma rede neural DenseNet-121. O classificador aprende a prever patologias com base no conjunto de dados combinado. b) DenseNet-121 é usado para classificar imagens com determinadas patologias em relação a outras imagens com diferentes patologias.

Em resumo, este processo aproveita imagens de raios-x sintéticas do tórax geradas por GAN para lidar com o desbalanceamento de classe nos dados de treinamento,

melhorando, em última análise, o desempenho do classificador de imagens de radiografia de tórax. Abaixo, Figura 2.7 segue a apresentação das imagens reais (a esquerda) e das amostras geradas (a direita) utilizando o método proposto.

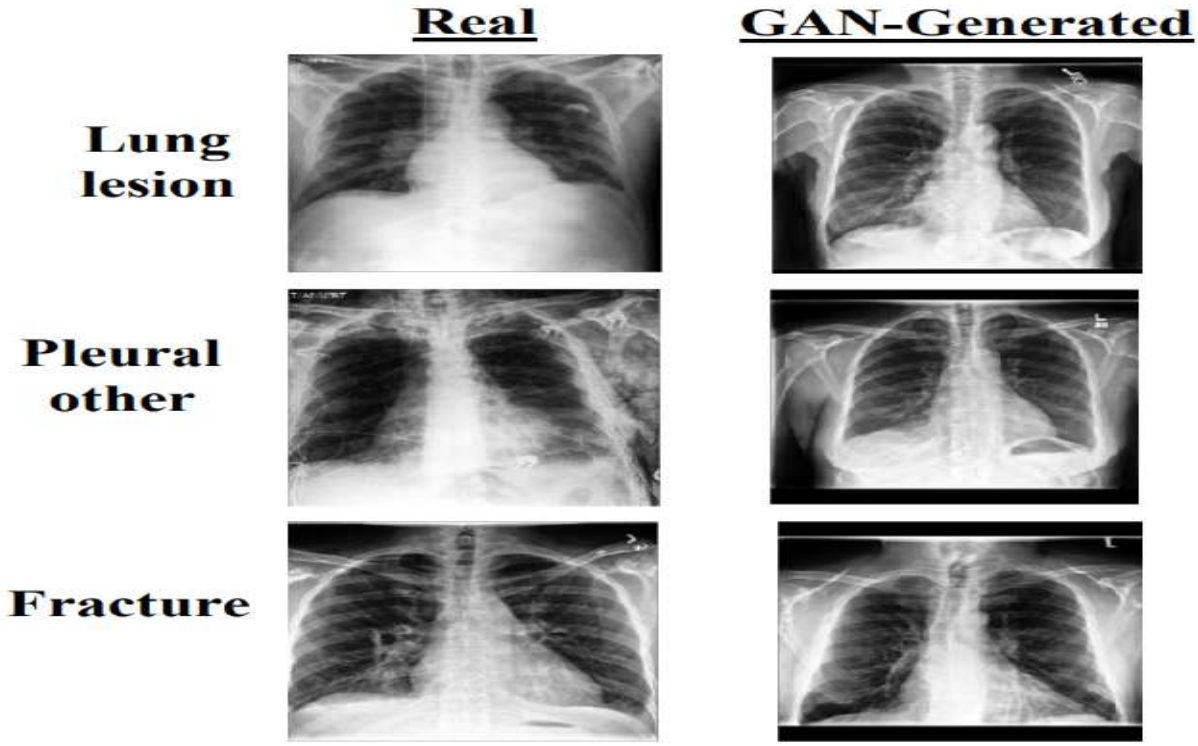


Figura 2.7: Amostra de radiografias de tórax sintéticas geradas por uma GAN pré-treinada para classes sub-representadas no CheXpert [96].

Motamed et al. [73], propuseram uma arquitetura de GAN para aumento de dados de radiografias de tórax para detecção semi-supervisionada de pneumonia e COVID-19, usando modelos generativos. O modelo proposto, *Inception Augmentation GAN* (IAGAN), foi inspirada no modelo *Data Augmentation Generative Adversarial Networks* (DAGAN) [4] utilizada para geração de dados sintéticos visando aumentar conjunto de dados para treinamento de outros modelos.

A Figura 2.8 apresenta a arquitetura do gerador do modelo IAGAN. Em cada iteração, o gerador recebe um vetor de ruído gaussiano z_i e um lote de imagens reais de treinamento x_i. Primeiramente, as imagens de entrada x_i são codificadas usando convolução e camadas de atenção para obter uma representação de baixa dimensão. Em seguida, essa representação é concatenada com o vetor de ruído z_i (após passar por uma camada densa e não linear). Essa abordagem visa não apenas utilizar a representação completa da imagem no discriminador, mas também obter uma representação de menor dimensão das imagens geradas pelo gerador, melhorando sua capacidade de generalização. A entrada dupla permite que o gerador treinado utilize imagens de diferentes classes e gere uma variedade mais ampla de imagens, enriquecendo o conjunto de dados de treinamento.

O uso de camadas de atenção em GANs, conforme mostrado na Figura 2.9, permite capturar uma gama mais ampla de recursos dentro da imagem. Enquanto as camadas de convolução simples focam em recursos locais restritos pelo seu campo receptivo, a camada de atenção usa três convoluções 1 × 1. Essa convolução 1 × 1 ajuda a reduzir o número de canais na rede. Duas das saídas da convolução, como ilustrado na Figura 2.9, são multiplicadas (por meio de multiplicação de matrizes) e o resultado é usado na produção do mapa de atenção.

O mapa de atenção atua como a probabilidade de cada píxel afetar a saída da terceira camada de convolução. Além disso, alimentar uma representação de baixa dimensão de uma imagem de entrada x permite que o gerador treinado utilize imagens de diferentes classes para produzir imagens semelhantes, nunca antes vistas, da classe em que foi treinado.

O uso das arquiteturas *Inception* e *Residual* aumenta a capacidade da GAN de capturar mais detalhes do espaço-imagem de treinamento, sem perder informações espaciais após cada camada de convolução e *pooling*. No entanto, é importante observar que tornar a rede geradora (G) mais profunda é teoricamente uma maneira de capturar mais detalhes na imagem, porém GANs profundos são instáveis e difíceis de treinar.

O discriminador (D) mostrado na Figura 2.10 é uma CNN de 4 camadas que mapeia uma imagem 2D para uma saída escalar que pode ser interpretada como a probabilidade da entrada fornecida ser uma imagem real de raios-x de tórax de dados de amostra (dados de treinamento) ou imagem G(z) gerada pelo gerador G.

A Figura 2.11 mostra a saída do gerador nos estágios inicial, intermediário e posterior (da esquerda para a direita, respectivamente) do treinamento nos conjuntos de dados saudáveis (Normal) e não saudáveis (com Pneumonia).

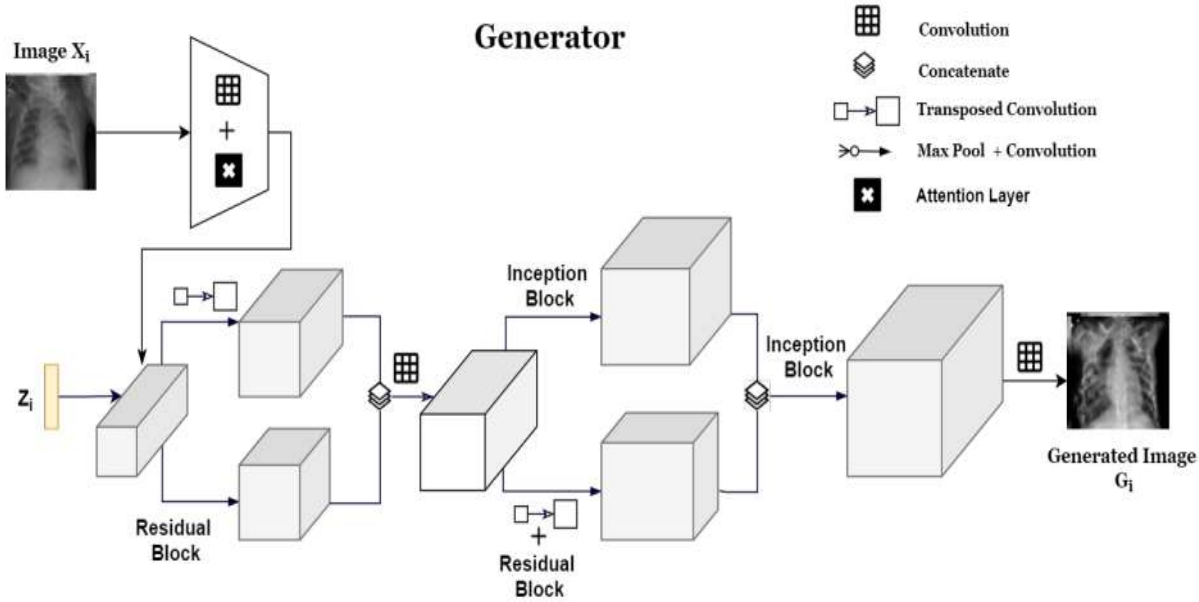


Figura 2.8: Arquitetura do Gerador IAGAN [73].

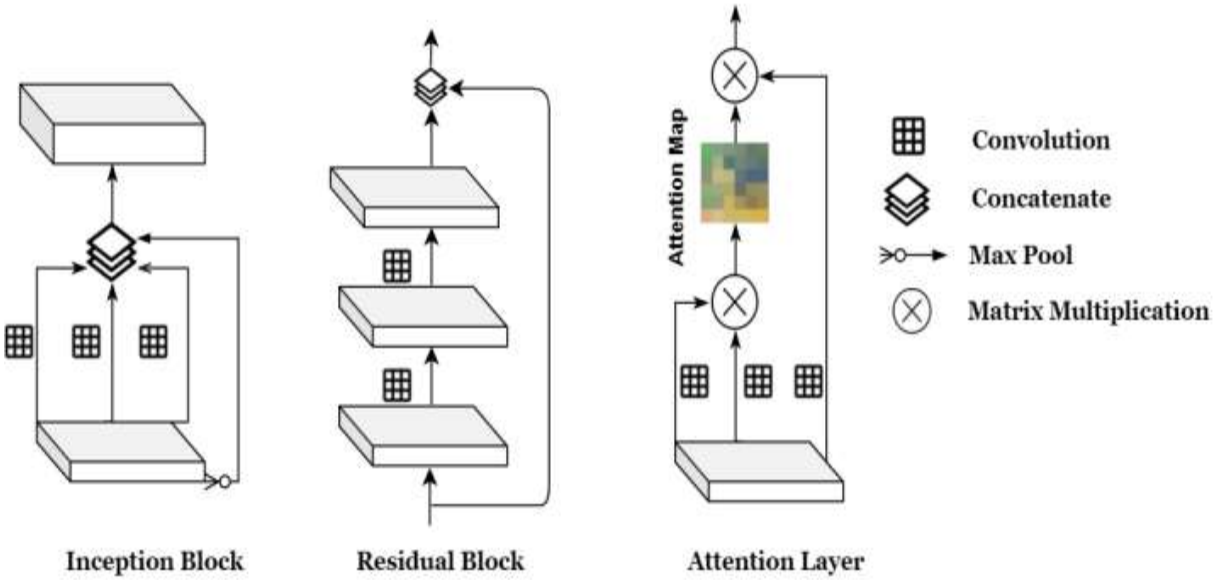


Figura 2.9: Detalhes da arquitetura do gerador do IAGAN [73].

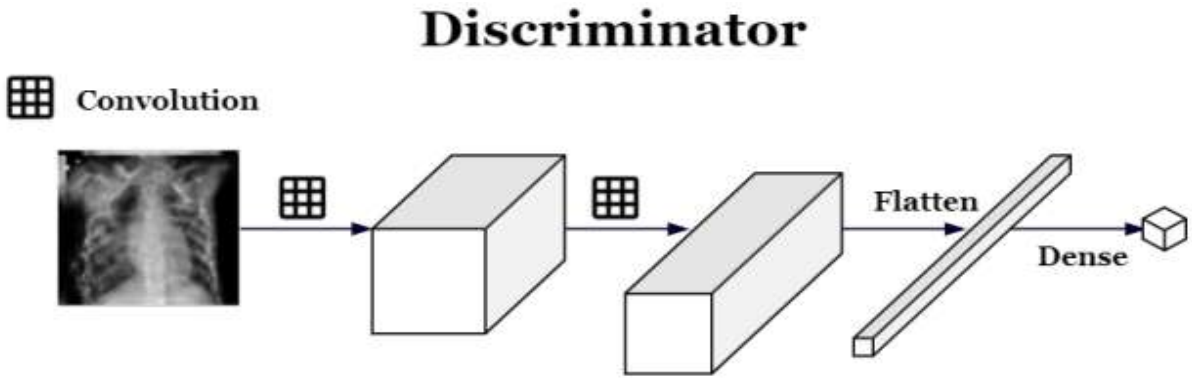


Figura 2.10: Arquitetura do Discriminador [73].

Mamadi et al. [70], investigaram o uso de GANs para a geração de imagens de radiografia de tórax para aumentar o conjunto de dados. O conjunto de dados aumentado é então usado para treinar a rede neural convolucional a classificar imagens em busca de anormalidades cardiovasculares. Segundo os autores, a abordagem de aumento de dados proposta mostrou maior precisão para classificação de radiografias de tórax, comparado a estratégia de aumento de dados tradicional. O trabalho proposto pelos autores revela uma importante contribuição, uma vez que remete a temas como desafios da escassez de dados médicos causados por questões de privacidade de informações sensíveis dos pacientes.

Em seu trabalho intitulado *Data augmentation using generative adversarial networks (gans) for gan-based detection of pneumonia and covid-19 in chest x-ray images*, Motamed et al. [73], propuseram uma arquitetura baseada em GANs para o aumento de dados de radiografias de tórax para detecção semi-supervisionada de pneumonia e COVID-19. Segundo os autores, a arquitetura proposta, Inception-Augmentation GAN (IAGAN), inspi-

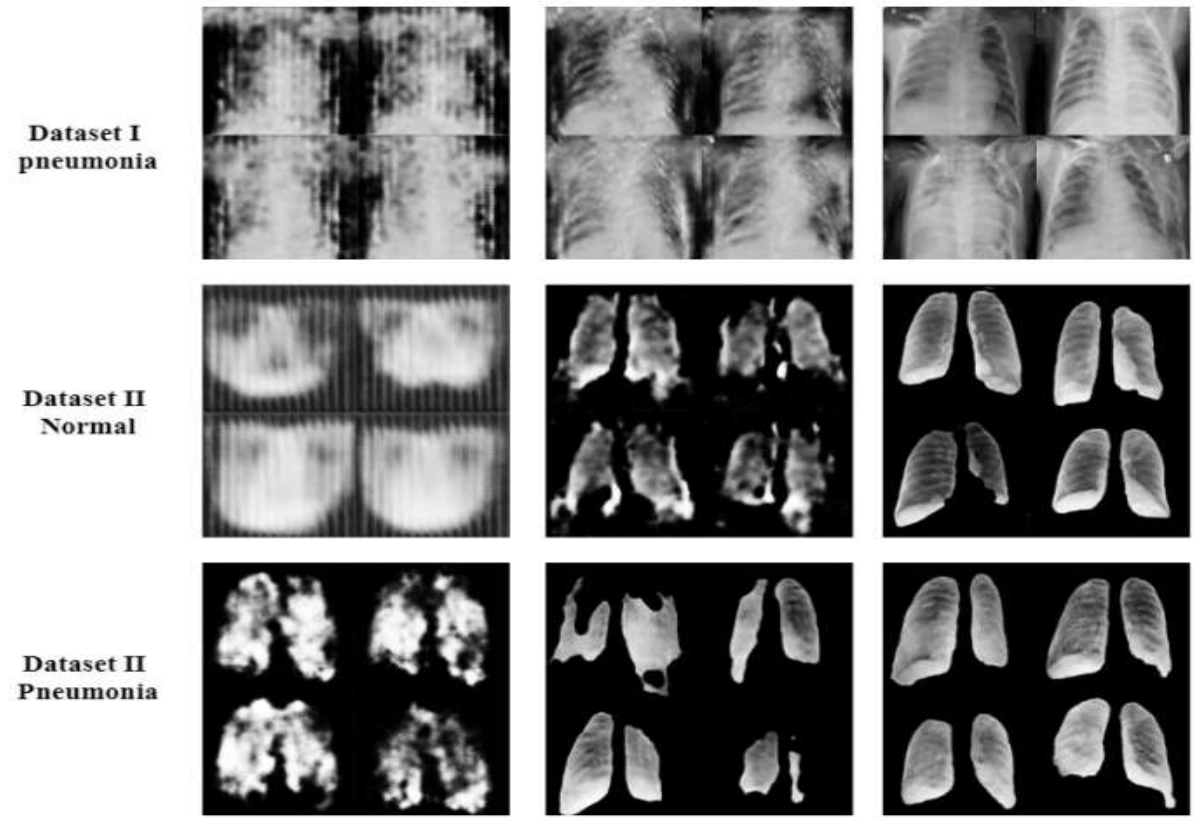


Figura 2.11: Imagens de raios-x do tórax geradas durante o treinamento [73].

rada no modelo *Data Augmentation Generative Adversarial Networks* (DAGAN) [4], pode ser usada para aumentar o conjunto de dados, melhorar a precisão da classificação de doenças como pneumonia e COVID-19 em radiografias de tórax e, principalmente, para melhorar o desempenho de outras arquiteturas de GANs.

2.2.4 *Diffusion Probabilistic Models (DPMs)*

Os modelos probabilísticos de difusão são classes poderosas de modelos generativos probabilísticos usados para aprender distribuições de dados complexas. Esses modelos conseguem isso utilizando dois estágios principais: o processo de difusão direta e o processo de difusão reversa [53].

O processo de difusão direta adiciona ruído aos dados de entrada, aumentando gradativamente o nível de ruído até que os dados sejam transformados completamente em ruído gaussiano. Este processo perturba sistematicamente a estrutura da distribuição dos dados. O processo de difusão reversa, também conhecido como processo de remoção de ruído, é então aplicado para recuperar a estrutura original dos dados a partir da distribuição de dados perturbada. Este processo desfaz efetivamente a degradação causada pelo processo de difusão direta. O resultado é um modelo generativo altamente flexível

e tratável que pode modelar com precisão distribuições complexas de dados a partir de ruído aleatório [53].

Diffusion Probabilistic Models (DPMs) também denominado simplesmente de *Diffusion Models* (DMs), são definidos como cadeias de Markov parametrizadas treinadas usando inferência variacional para produzir amostras correspondentes aos dados após um tempo finito. As transições dessas cadeias são aprendidas para reverter o processo de difusão, que é uma cadeia de Markov que gradualmente adiciona ruído aos dados na direção oposta da amostragem até que o sinal seja totalmente destruído. Esses modelos são poderosos na estimativa de densidade e na qualidade da amostra, mas sua operação no espaço de pixels, adicionando ou removendo ruído a um tensor do mesmo tamanho da imagem original, resulta em baixa velocidade de inferência e alto custo computacional [40].

Outro aspecto importante a ser destacado é a capacidade dos modelos probabilísticos de difusão de gerar uma variedade maior de imagens, demonstrando não sofrer o problema de *Mode Collapse*, comumente enfrentado pelos modelos GANs. Isso se deve à capacidade dos DMs de preservar a estrutura semântica dos dados. No entanto, esses modelos demandam alta capacidade computacional, e o treinamento requer uma memória muito grande, o que pode ser inviável experimentar o método. Além disso, o tempo de treinamento dos DMs é muito alto, podendo durar dias ou até meses. Isto porque esses modelos tendem a ficar presos nas complexidades imperceptíveis nas imagens [40, 93].

Diffusion Probabilistic Model (DPM) tem sido usado amplamente no campo da visão computacional. Sua aplicação na geração de imagens tem demonstrado uma impressionante capacidade generativa, despertando ampla discussão na comunidade científica. Inspirados pelo sucesso do DPM, Wu et al. [108], propuseram Medsegdiff, um modelo baseado em DPM para tarefas de segmentação de imagens médicas.

Harder et al. [55], no trabalho intitulado *Denoising diffusion probabilistic models for 3D medical image generation*, mostraram que modelos probabilísticos de difusão podem ser utilizados na sintetização de dados médicos de alta qualidade para ressonância magnética (MRI) e tomografia computadorizada (TC).

2.2.5 *Stable Diffusion Model*

Em junho de 2021, a OpenAI publicou um artigo intitulado *Diffusion Models Beat GANs on Image Synthesis* [27]. Segundo os autores envolvidos, *Stable Diffusion Models* (SDMs) são capazes de produzir imagens de alta qualidade, oferecendo propriedades desejáveis, como fácil escalabilidade. Ainda segundo os autores, esses modelos geram amostras removendo gradualmente o ruído de um sinal. Uma versão desses modelos

difusos é o *Latent Diffusion Model*, que, na verdade, é o algoritmo por trás da *Stable Diffusion Model*.

2.2.6 *Latent Diffusion Models (LDMs)*

Esta seção traz contribuições de diferentes autores sobre o tema *Latent Diffusion Models*. Inicialmente, apresentamos o conceito por trás dos LDMs, seguido da discussão dos trabalhos correlatos.

Os LDMs são modelos generativos capazes de aproveitar o poder perceptivo dos modelos GANs, a capacidade de preservação detalhada dos *Diffusion Models* e a capacidade semântica dos Transformers (são capazes de capturar a estrutura semântica em textos e em imagens, mas eles exigem uma grande quantidade de dados para o treinamento), fundindo todos os três em um modelo [89]. Os LDMs foram apresentados por Rombach et al. [89], no artigo intitulado *High-Resolution Image Synthesis with Latent Diffusion Models*. Eles são modelos *text-to-image* capazes de, dado *inputs* de textos (*prompts*), gerar imagens fotorealistas. São constituídos por dois processos fundamentais: ***Forward Process*** ou ***Diffusion Process*** (Figura 2.12) e ***Reverse Process*** ou ***Denoising Process*** (Figura 2.13). Eles operam com base no princípio de que, ao invés de aplicar o processo de difusão diretamente na imagem de entrada de alta dimensão, a entrada é projetada em um espaço latente de menor dimensão onde a difusão (inserção de ruídos aleatórios) é aplicada. Para exemplificar os dois processos, utilizamos as imagens do *dataset* do Programa de Controle de Tuberculose do Departamento de Saúde e Serviços Humanos do Condado de Montgomery [45].

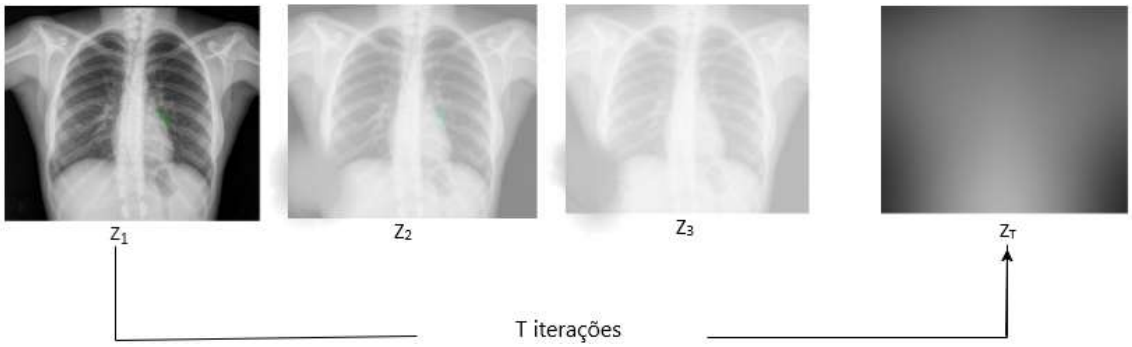


Figura 2.12: *Forward Process* ou *Diffusion Process* [45].

No processo de difusão (*Diffusion Process*), a imagem é corrompida pela adição gradual de ruído até que a imagem se torne um ruído aleatório por completo (até a imagem ficar irreconhecível). Este é o processo que gera dados necessários para treinar (ensinar) o modelo em como remover ruídos das imagens (no caso é o modelo *Denoising U-Net* responsável pela remoção de ruídos). Para treinar a rede U-Net, começa-se com uma imagem (por exemplo, a imagem mostrada na Figura 2.13 representada por Z_7) e

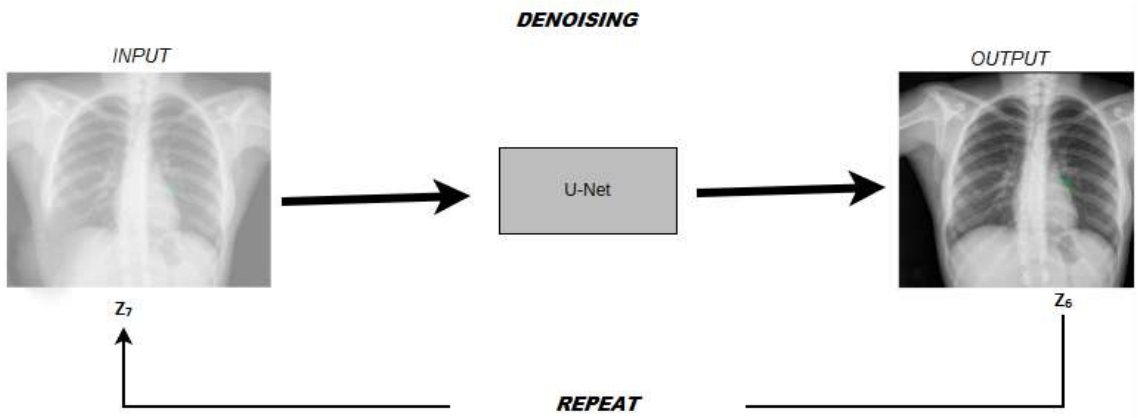


Figura 2.13: *Reverse Process* ou *Denoising Process* [45].

pede-se para que a U-Net aprenda a obter a imagem anterior (representada por Z_6, Figura 2.13) na mesma sequência. Basicamente, a U-Net remove uma camada de ruído da imagem de entrada a cada iteração. Esse processo ocorre repetidamente, com cada par de imagem, até que o ruído seja completamente removido em cada um desses pares de imagem. Em suma, no processo inverso, uma série de cadeias de Markov são usadas para recuperar os dados do ruído Gaussiano, removendo gradualmente o ruído previsto a cada intervalo de tempo.

O trabalho publicado no artigo *High-Resolution Image Synthesis with Latent Diffusion Models* foi logo tida como base para o desenvolvimento de importantes outros trabalhos como os abordados na sequência:

Pinaya et al. [82], exploraram o uso de *Latent Diffusion Models* para gerar imagens sintéticas a partir de imagens cerebrais de alta resolução. Os autores usaram imagens de ressonância magnética do conjunto de dados biomédico do UK Biobank para treinar seus modelos para aprender sobre a distribuição probabilística de imagens cerebrais, condicionadas a covariáveis, como idade, sexo e volumes de estrutura cerebral. A motivação por trás desse trabalho veio da escassez do conjunto de dados médicos para atender as demandas de dados médicos em projetos de imagiologia médica (em inglês *medical imaging)*.

Nos experimentos realizados, os autores utilizaram LDMs, que combinam o uso de *autoencoders* para compactar os dados de entrada em uma representação latente de menor dimensão, incorporando as propriedades de modelagem generativa dos modelos de difusão. O modelo de compressão desempenha um papel essencial ao permitir a escalabilidade de imagens médicas de alta resolução. O *autoencoder* foi treinado com uma combinação da perda perceptiva e um objetivo adversário baseado em *patch* (Aqui, *patch* refere-se a um pequeno trecho ou área específica de uma imagem, usado para modificar ou manipular partes específicas da imagem visando criar adversários ou enganar algoritmos de aprendizado de máquina.). Além disso, aplicou-se uma regularização Kullback-Leibler (KL) ao espaço latente, atuando como um regularizador para controlar o

grau de compressão nesse espaço. O codificador mapeia a imagem do cérebro para uma representação latente com dimensões de 20 × 28 × 20 [82].

Após o treinamento do modelo de compressão, as representações latentes do conjunto de treinamento são utilizadas como entrada para o modelo de difusão. Esses modelos foram condicionados com base em características como idade, sexo, volume ventricular e volume cerebral em relação ao volume intracraniano. Para realizar esse condicionamento, os autores adotaram uma abordagem híbrida, combinando a concatenação das informações de condicionamento com os dados de entrada e a utilização de mecanismos de atenção cruzada [82]. Na Figura 2.14 apresentamos as imagens de ressonância magnética da cabeça reais (na parte de cima) e as amostras geradas utilizando LDM (na parte de baixo).

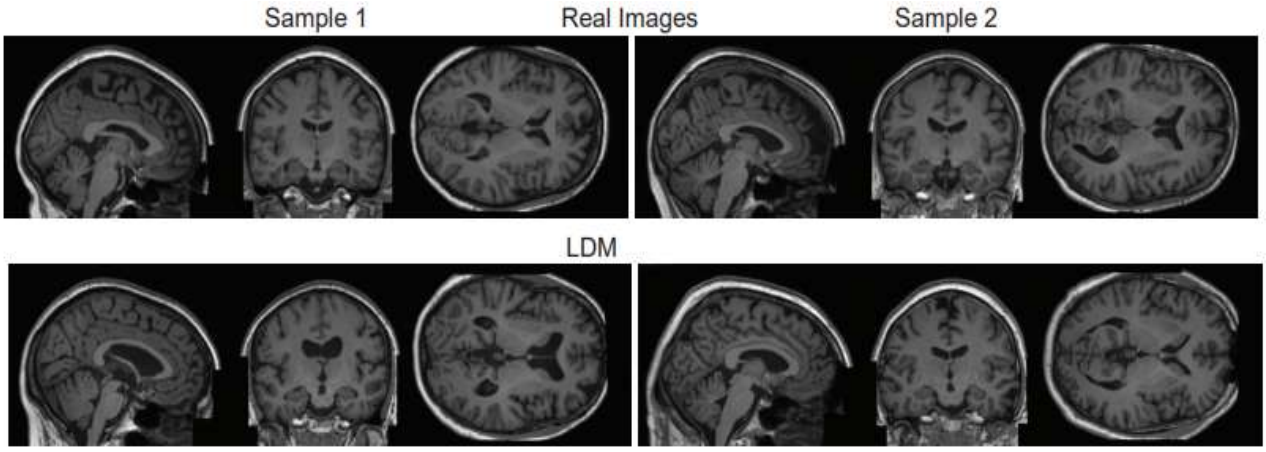


Figura 2.14: Amostras reais e sintéticas de ressonância magnética da cabeça geradas usando LDM [82].

Ali et al. [2] exploraram a síntese de imagens médicas utilizando modelos de difusão neural. Como abordagem, eles primeiro empregaram um modelo DALLE2 pré-treinado para gerar imagens de raios-x e tomografia computadorizada dos pulmões a partir de um prompt de texto de entrada. Em seguida, treinaram um modelo de difusão estável com um conjunto de 3.165 imagens de raios-x em um servidor equipado com a GPU NVIDIA Quadro RTX 8000, que possui 48 GB de memória. Como procedimento de avaliação, realizaram uma análise qualitativa, na qual dois radiologistas independentes rotularam amostras escolhidas aleatoriamente como reais, falsas ou incertas. Segundo os autores, os resultados demonstraram que as imagens geradas pelo modelo de difusão podem traduzir características que, de outra forma, seriam muito específicas de determinadas condições médicas em imagens de radiografia de tórax ou tomografia computadorizada. Abaixo, apresentamos amostras das imagens de raios-x dos pulmões (Figura 2.15) e amostras das imagens de tomografia computadorizada de pulmões (Figura 2.16) geradas com o modelo de difusão.

Figura 2.15: Amostras de imagens de raios-x dos pulmões geradas com o modelo de difusão [2].

Figura 2.16: Amostras de imagens de tomografia computadorizada de pulmões geradas com o modelo de difusão [2].

Packhauser et al. [78] utilizaram o LDM [89] para gerar imagens de radiografias de tórax de alta qualidade, ao mesmo tempo que propunham uma estratégia de amostragem para preservar a privacidade das informações biométricas sensíveis durante o processo de geração. O método consiste em aplicar o LDM proposto por Rombach et al. [89], aproveitando *autoencoders* pré-treinados. Nessa abordagem, as imagens de entrada são incorporadas a um espaço latente de tamanho 64 × 64 × 3 usando o codificador do *autoencoder* VQ-VAE com 32, 64 e 128 canais em cada estágio. Além disso, um modelo de difusão opera nesse espaço inferior, dominado por frequências mais baixas, executando 1000 etapas de remoção de ruído com uma U-Net (32, 128 e 256 canais). Na etapa final, o decodificador do *autoencoder* aumenta a resolução espacial para 256×256 píxeis, introduzindo frequências mais altas. As informações condicionais são incorporadas usando uma tabela de consulta treinável, combinando os *embeddings* com o processo de difusão através de atenção cruzada no gargalo da U-Net.

Para treinar o modelo, os pesquisadores utilizaram um conjunto de dados composto por radiografias de tórax de 30.805 pacientes. Os metadados fornecem informações sobre 14 rótulos de anormalidades correspondentes, incluindo Atelectasia, Cardiomegalia, Consolidação, Edema, Efusão, Enfisema, Fibrose, Hérnia, Infiltração, Massa, Nódulo, Espessamento Pleural, Pneumonia e Pneumotórax. Indivíduos saudáveis são rotulados com uma classe adicional, indicando a ausência dessas anomalias.

A avaliação da utilidade do conjunto de dados gerado envolveu a análise das imagens em uma tarefa de classificação de anormalidades torácicas. Segundo os autores, a abordagem proposta superou os métodos baseados em GAN. Abaixo, você pode conferir amostras de imagens de raios-x geradas (Figura 2.17).

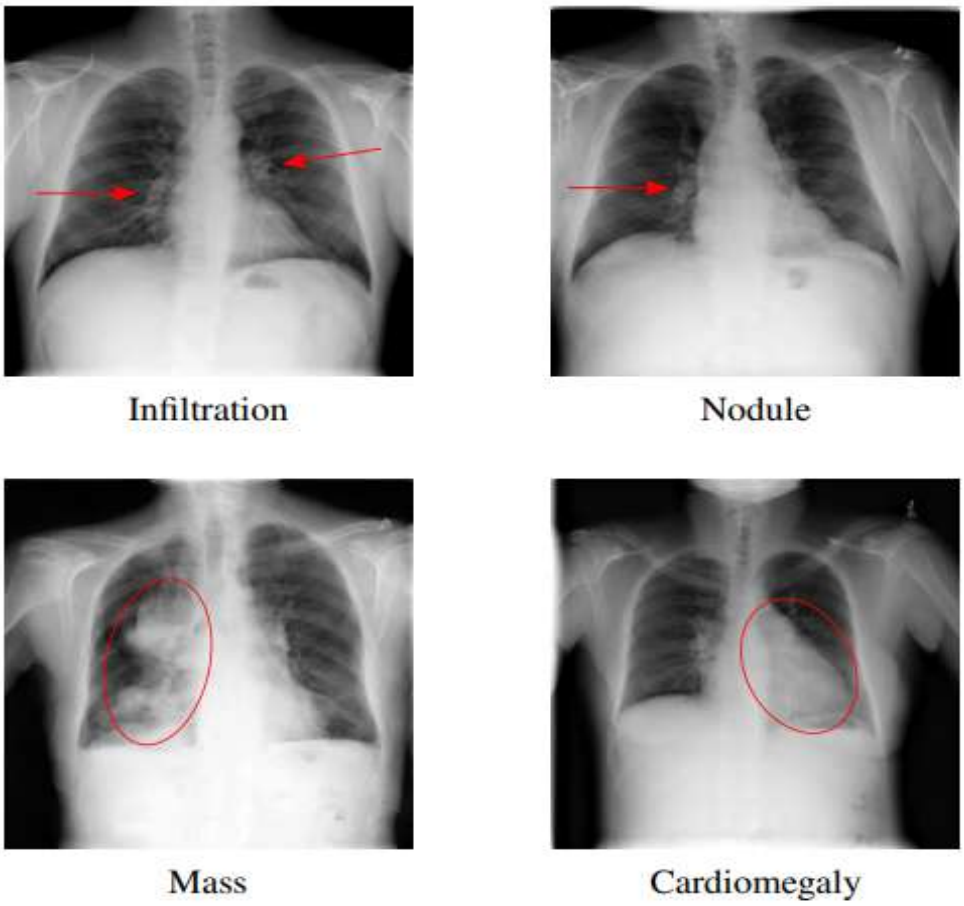


Figura 2.17: Imagens geradas por LDM e selecionadas aleatoriamente. As posições e áreas de anormalidades induzidas são destacadas com setas ou círculos vermelhos [78].

Nos modelos de síntese de ressonância magnética baseados em difusão, é comum empregar a abordagem unimodal. No entanto, visto que esses modelos operam no domínio da imagem original, eles exigem muita memória e são menos viáveis para síntese multimodal. Adicionalmente, frequentemente não são capazes de preservar a estrutura anatômica na ressonância magnética. Para diminuir o impacto deste problema, Jiang et al. [48], propuseram o modelo de síntese de ressonância magnética multimodal baseado em difusão, denominado modelo de difusão latente condicionada (em inglês *CoLa-Diff: Conditional Latent Diffusion Model for Multi-Modal MRI Synthesis*). Nesse modelo, eles propuseram uma arquitetura projetada para reduzir o consumo de memória operando no espaço latente. Para resolver problemas relacionados a compressão e ruído presentes no espaço latente, os autores utilizaram uma abordagem de filtragem cooperativa (Técnica que visa reduzir o ruído de alta dimensão gerado durante o processo de síntese, que de outra forma poderia impactar negativamente a qualidade da geração de

imagens multimodais.) inspirada em técnicas de filtragem colaborativa. Além disso, para garantir a preservação das estruturas anatômicas, consideram a inclusão de máscaras de regiões cerebrais como prioritárias para distribuições de densidade para orientar o processo de difusão. E para aproveitar as informações multimodais de forma eficaz, eles implementaram uma técnica de adaptação automática de peso. Conforme os autores, os experimentos demonstraram que o método CoLa-Diff proposto supera outros métodos de síntese de ressonância magnética do estado da arte (até o momento da apresentação), indicando que o CoLa-Diff tem um potencial de ser uma ferramenta eficaz para facilitar a síntese multimodal de ressonância magnética [53, 48].

O modelo proposto foi avaliado em dois conjuntos de dados de ressonância magnética cerebral com múltiplos contrastes: o BRATS 2018 e o IXI. O BRATS 2018 contém exames de ressonância magnética de 285 pacientes com glioma, enquanto o conjunto de dados IXI consiste em 200 ressonâncias magnéticas multicontraste de cérebros saudáveis. A seguir, apresentamos amostras geradas após o treinamento do modelo. O modelo foi treinado em duas NVIDIA RTX A5000, 24 GB de memória, utilizando o otimizador Adam no *framework* PyTorch. Além disso, o modelo foi avaliado em dois conjuntos de dados de ressonância magnética cerebral com múltiplos contrastes: o BRATS 2018 e o IXI. O BRATS 2018 contém exames de ressonância magnética de 285 pacientes com glioma, enquanto o conjunto de dados IXI consiste em 200 ressonâncias magnéticas multicontraste de cérebros saudáveis. A seguir, apresentamos amostras geradas após o treinamento do modelo (Figura 2.18).

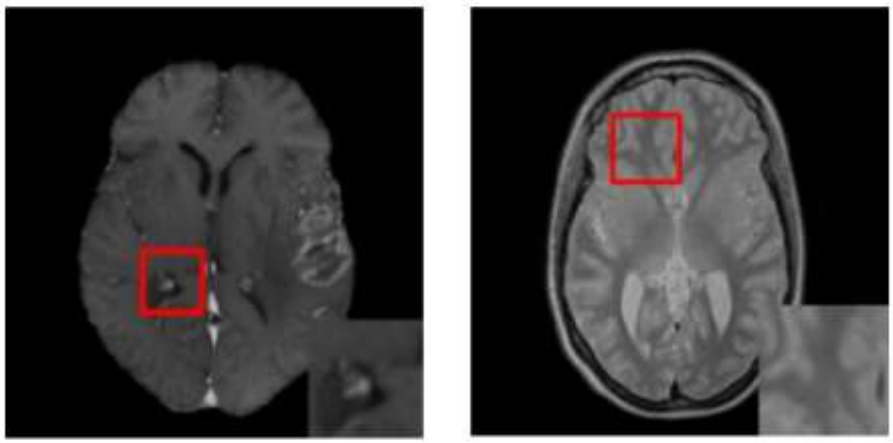


Figura 2.18: Amostras de imagens de ressonância magnética do cérebro geradas com o modelo CoLa-Diff [48].

Para mitigar a escassez de modelos de geração de imagens que representem fielmente conceitos médicos e, ao mesmo tempo, forneçam diversidade composicional e a escassez existente de conjuntos de dados de imagens médicas anotadas e de alta qualidade, Chambon et al. [15], adaptaram *Latent Diffusion Model* pré-treinado em um corpus de radiografia de tórax, publicamente disponíveis e seus correspondentes relatórios radiológicos, e desenvolveram o modelo RoentGen. Segundo os autores, RoentGen é um

modelo generativo capaz de sintetizar imagens de raios-x de tórax de alta fidelidade e é capaz de inserir, combinar e modificar as aparências de imagens de vários achados de raios-x de tórax por meio de *prompts* de texto em linguagem médica de formato livre.

Para desenvolver o modelo generativo capaz de incorporar uma variedade de conceitos médicos formulados em linguagem natural para o domínio de raios-x do tórax, os autores usaram o conjunto de dados MIMIC-CXR, que está disponível publicamente. Esse conjunto de dados contém 377110 imagens e seus relatórios radiológicos associados. Abaixo, apresentamos algumas amostras geradas pelo modelo (Figura 2.19).

Figura 2.19: Amostras de imagens sintéticas radiografias torácicas geradas. As radiografias apresentam altos níveis de detalhes: edema (canto superior direito), nebulosidade peri-hilar (pontas de setas brancas) e manguito peribrônquico (ponta de seta preta), ambas características observadas no edema pulmonar, podem ser observadas. Para pneumotórax (linha inferior, terceira imagem a partir da esquerda), uma linha fina representando o revestimento pleural visceral do pulmão parcialmente colapsado pode ser delineada (linha tracejada) [48].

Os trabalhos apresentados acima foram de fundamental importância para o entendimento e utilização dos Modelos de Difusão Latente para a geração de dados sintéticos fotorrealistas propostos neste trabalho. Embora todos tenham auxiliado na compreensão dos LDMs na perspectiva proposta, os trabalhos de autores como Chambon et al. [15]; Jiang et al. [48]; Packhauser et al. [78] e Pinaya et al. [82] são os que mais se destacam em termos de geração de imagens médicas, uma vez que abordam os cuidados que devem ser tomados quanto à proteção de privacidade, os desafios da escassez de dados, a relevância clínica e adaptação de domínio no campo das imagens médicas. Além disso, esses trabalhos mostraram a possibilidade e o potencial de uso de dados médicos sintéticos para diferentes finalidades, como para aumentar o conjunto de dados reais e, dessa forma, melhorar o desempenho de algoritmos de análise de imagens médicas. O próximo capítulo descreve a metodologia desenvolvida neste trabalho.

3. MÉTODO PROPOSTO

Este capítulo é dedicado a apresentação do método proposto para o desenvolvimento deste trabalho e está dividido em três seções principais. A Seção 3.1 é destinada à apresentação da visão genérica das arquiteturas do treinamento dos modelos e de geração das imagens. A Seção 3.2, consiste na apresentação de informações do *dataset* utilizado. Por fim, Seção 3.3, apresenta a estrutura e treinamento de *Latent Diffusion Models* (LDM). Além disso, esta segunda seção está subdividida em subseções que abordam a metodologia utilizada para o treinamento do modelo e para a geração de imagens.

3.1 *Visão geral da arquitetura de treinamento dos modelos e de geração das imagens*

As Figuras 3.1 e 3.2 apresentam, de forma resumida, a visão geral da arquitetura do treinamento do modelo e a arquitetura de geração das imagens de raios-x do tórax. Posteriormente, na Figura 3.8, detalhamos o fluxo de geração dessas imagens. Na Figura 3.1, é apresentado o *Data input*, que representa o *dataset* utilizado para treinar o *Stable Diffusion Foundation Model* (SDFM). Cada imagem é acompanhada de seu relatório clínico, representado na figura por um arquivo de texto (txt). Para o treinamento do SDFM, é utilizado a interface Kohya-ss. Após o treinamento, os modelos treinados são gerados como saída. Essa saída, por sua vez, é utilizada como entrada no *software* AUTOMATIC1111 (interface de geração) para gerar imagens de raios-x do tórax (Figura 3.2).

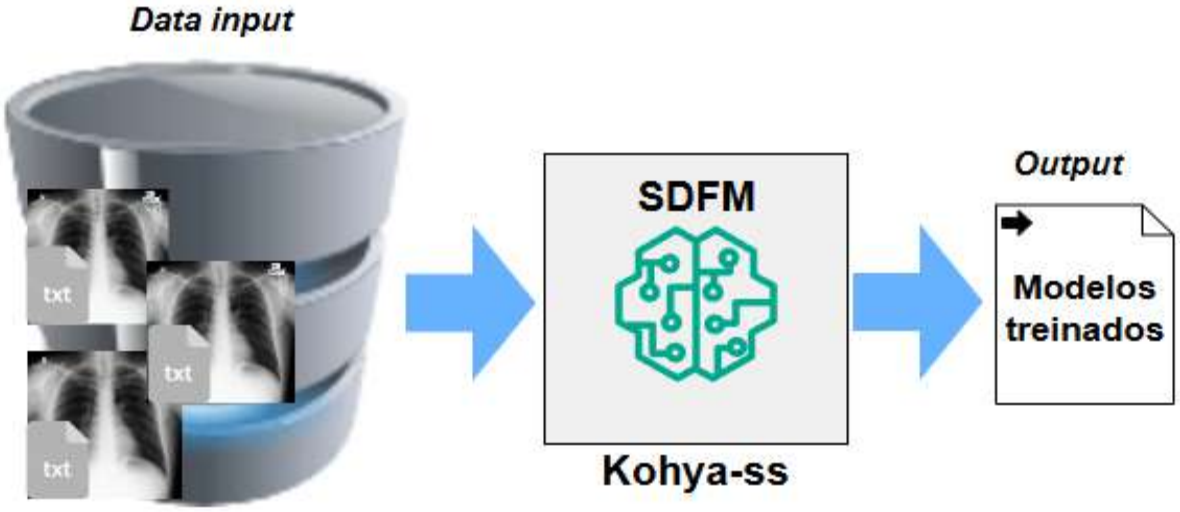


Figura 3.1: Visão genérica da arquitetura de treinamento dos modelos.

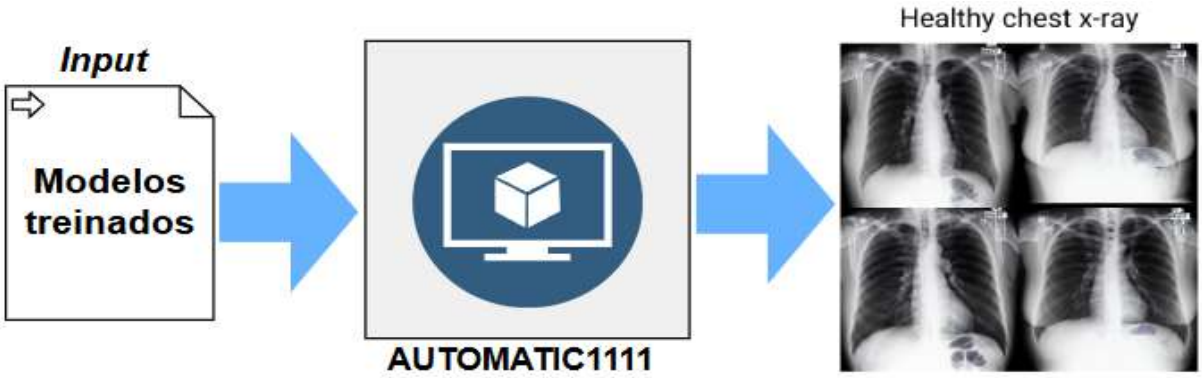


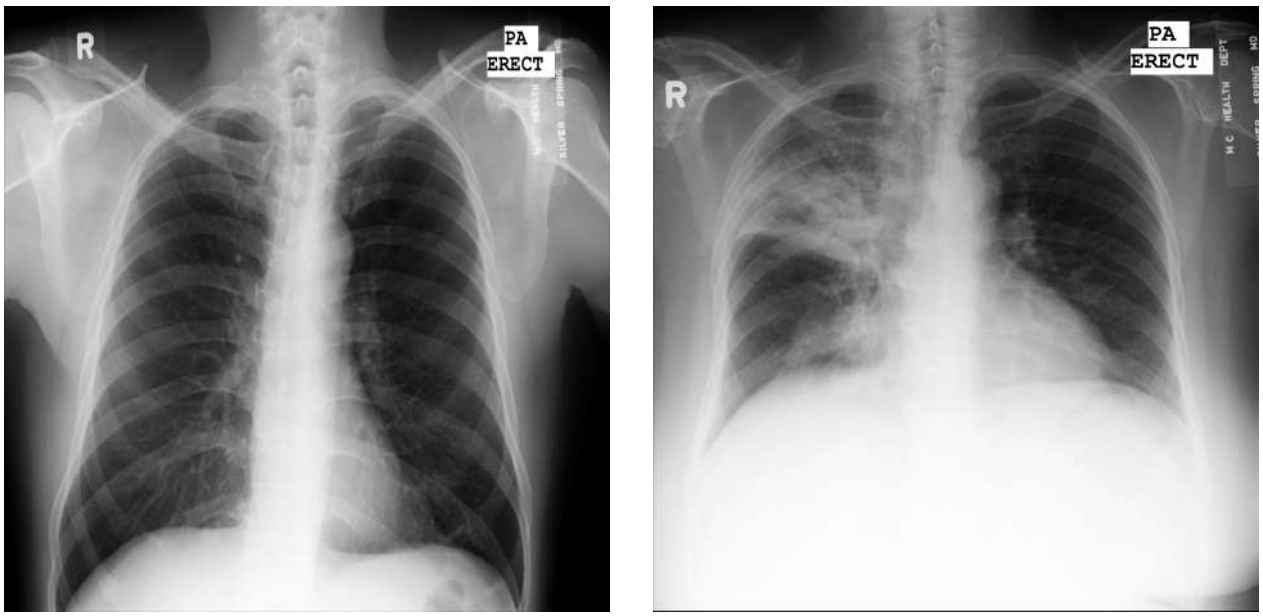
Figura 3.2: Visão genérica da arquitetura de geração das imagens de raios-x do tórax.

3.2 *Datasets*

Neste trabalho, utilizamos o conjunto de dados do banco de imagens digitais para tuberculose, criado pela *National Library of Medicine* em colaboração com o *Department of Health and Human Services*, Montgomery County, Maryland, USA (Programa de Controle de Tuberculose do Departamento de Saúde e Serviços Humanos do Condado de Montgomery, Maryland, EUA). O conjunto de dados, disponível publicamente, contém 138 radiografias póstero-anteriores de tórax, das quais 80 são de casos normais e 58 são de casos anormais com manifestações consistentes de tuberculose (Figura 3.3a e Figura 3.3b). Todas as imagens são anonimizadas e estão disponíveis em formato PNG. Além disso, o conjunto de dados inclui anotações de consenso de duas radiologistas para imagens redimensionadas de 1024 × 1024 e um relatório que descreve os resultados do exame de imagem [45].

Com base nas 138 radiografias disponíveis, utilizamos um conjunto composto por 30 imagens (50% de imagens saudáveis e 50% de imagens não saudáveis) para treinar os modelos na primeira fase do nosso estudo. Optamos por trabalhar com um conjunto de dados deste tamanho na fase inicial para que pudéssemos obter uma orientação mais precisa sobre os próximos passos a serem seguidos.

Na segunda fase, preparamos sete *datasets* de 5, 10, 20, 40, 60, 80 e 100 imagens para treinar os modelos. Em cada conjunto, mantivemos uma proporção equilibrada, com 50% de imagens saudáveis e 50% de imagens não saudáveis, todas acompanhadas de seus respectivos relatórios clínicos. O objetivo era testar o impacto da quantidade de imagens no treino de *Foundation Models*.



(a) Imagem de raios-x normal.

(b) Raio-x anormal.

Figura 3.3: Texto descrito no laudo das imagens [45]: (a) *Normal chest x-ray*; (b) *Large infiltrate Right Upper Lobe with cavitation plus infiltrate in RML. Consistent with active cavitary TB.*

Na sequência, realizamos a tarefa de encontrar no *dataset* utilizado a ocorrência de principais termos que fazem parte dos *prompts* utilizados na geração de imagens. Esse processo, denominado Contagem de Frequência de Palavras ou Frequência de Termos, é um procedimento comum em Processamento de Linguagem Natural (PNL) e análise de texto que visa identificar as palavras usadas com mais frequência. Para o nosso trabalho, isso é útil, uma vez que possibilita fazer avaliação das imagens geradas, com base nos *prompts* existentes nos dados de treino. As análises feitas incluem as imagens geradas pelos *prompts* que têm ou não termos que ocorreram no *dataset* de treino.

Antes de procedermos ao cálculo das ocorrências dos termos, realizamos a preparação dos nossos conjuntos de dados, ou *datasets*, que foram usados para treinar nossos modelos. Nesse processo, desconsideramos palavras comuns em inglês, conhecidas como *stopwords*, além de pontuações. Também não fizemos distinção entre letras maiúsculas e minúsculas.

Nas Tabelas 3.1, 3.2 e 3.3, apresentamos os 100 termos (80 termos no caso do conjunto de 20 imagens) que mais ocorreram nos conjuntos de dados de 30 imagens, utilizados para treinar os modelos na primeira fase e nos conjuntos de 20,e 100 imagens. Estes últimos, fazem parte dos sete conjuntos utilizados para treinar os modelos na segunda fase.

Os experimentos realizados foram organizados em duas fases. Na primeira, utilizamos um conjunto de dados composto por 30 imagens para o treinamento dos modelos. Na segunda fase, expandimos a abordagem e utilizamos sete conjuntos de dados para

o mesmo propósito. Esses conjuntos continham 5, 10, 20, 40, 60, 80 e 100 imagens, respectivamente.

É importante ressaltar que esses conjuntos são cumulativos, ou seja, o conjunto de 5 imagens está contido no de 20, que por sua vez está contido no de 40, e assim por diante, até o conjunto de 100 imagens.

No entanto, optamos por destacar as ocorrências dos termos nos conjuntos de 20 e 100 imagens, pois esses são os conjuntos que, inicialmente, seriam submetidos à avaliação dos especialistas. A descrição detalhada dessas fases será apresentada mais adiante.

Os termos nas células vermelhas são termos (ou seus variantes) utilizados nos *prompts* de geração de imagens de raios-x do tórax. Na Tabela 3.4, apresentamos a ocorrência desses termos no *dataset* de 30 imagens, utilizado para treinar os modelos na primeira fase. Já na Tabela 3.5, apresentamos a ocorrência dos termos utilizados nos *prompts* de geração de imagens nos *datasets* compostos por 20 e 100 imagens, utilizados para treinar os modelos na segunda fase. É importante destacar que as imagens presentes no *dataset* de 20 imagens também estão incluídas no conjunto de *dataset* de 100 imagens. No entanto, essa mesma condição não se aplica necessariamente ao *dataset* de 30 imagens.

3.3 Estrutura e Treinamento de *Latent Diffusion Models* (LDM)

A abordagem proposta neste trabalho tem como base o trabalho apresentado por ROMBACH et al. [89], no artigo intitulado *High-Resolution Image Synthesis with Latent Diffusion Models*. Para exemplificar o comportamento do modelo LDM, vamos dividir o seu funcionamento em quatro etapas principais:

- A primeira etapa consiste na extração de uma representação mais compacta da imagem, utilizando o *encoder* ε localizado no canto superior esquerdo da Figura 3.4. Ao contrário de outros métodos, LDM funciona no espaço latente definido pelo *encoder*, e não no espaço de pixel [89, 114].

- Em seguida, o ruído gaussiano é adicionado à imagem (parte central superior da Figura 3.4, denominado *Diffusion Process*), como parte do processo de difusão que vai de z (representação compacta da imagem) a z_T (para o caso em que se tem T passos de adição de ruído).

- A representação z_T (representação compacta da imagem e com o ruído já adicionado nela) é então executada por uma Rede Neural Convolucional U-Net (localizada na parte central inferior da Figura 3.4). A Rede Neural Convolucional U-Net tem a função

Tabela 3.1: Os 80 termos mais frequentes no *dataset* de 20 imagens.

	Termos	Ocorrência	Porcentagem (%)		Termos	Ocorrência	Porcentagem (%)
1	*patients*	40	11,8343	42	*infoltrate*	1	0,2959
2	*sex*	20	5,9172	43	*outer*	1	0,2959
3	*age*	20	5,9172	44	*fibronodular*	1	0,2959
4	*years*	20	5,9172	45	*typical*	1	0,2959
5	*human*	20	5,9172	46	*appearance*	1	0,2959
6	*chest*	20	5,9172	47	*radiotherapy*	1	0,2959
7	*xray*	20	5,9172	48	*also*	1	0,2959
8	*right*	11	3,2544	49	*previous*	1	0,2959
9	*lobe*	10	2,9586	50	*cardiac*	1	0,2959
10	*tuberculosis*	10	2,9586	51	*surgery*	1	0,2959
11	*healthy*	10	2,9586	51	*suspicious*	1	0,2959
12	*normal*	10	2,9586	52	*extensive*	1	0,2959
13	*f*	9	2,6627	53	*bilaterally*	1	0,2959
14	*active*	8	2,3669	54	*cavity*	1	0,2959
15	*infiltrate*	7	2,0710	55	*moderate*	1	0,2959
16	*upper*	7	2,0710	56	*acidfast*	1	0,2959
17	*pleural*	7	2,0710	57	*bacilli*	1	0,2959
18	*effusion*	7	2,0710	58	*smears*	1	0,2959
19	*left*	5	1,4793	59	*ribonucleic*	1	0,2959
20	*cavitary*	4	1,1834	60	*acid*	1	0,2959
21	*large*	3	0,8876	61	*probes*	1	0,2959
22	*likely*	3	0,8876	62	*pos*	1	0,2959
23	*cavitation*	2	0,5917	63	*mycobacterium*	1	0,2959
24	*middle*	2	0,5917	64	*small*	1	0,2959
25	*consistent*	2	0,5917	65	*areas*	1	0,2959
26	*infiltrates*	2	0,5917	66	*iodoamphetamine*	1	0,2959
27	*lung*	2	0,5917	67	*positive*	1	0,2959
28	*nodular*	2	0,5917	68	*tuberculin*	1	0,2959
29	*lower*	2	0,5917	69	*skin*	1	0,2959
30	*findings*	2	0,5917	70	*test*	1	0,2959
31	*plus*	1	0,2959	71	*bilateral*	1	0,2959
32	*best*	1	0,2959	72	*miliary*	1	0,2959
33	*seen*	1	0,2959	73	*nodules*	1	0,2959
34	*lateral*	1	0,2959	74	*diffusely*	1	0,2959
35	*view*	1	0,2959	75	*old*	1	0,2959
36	*scar*	1	0,2959	76	*calcified*	1	0,2959
37	*soft*	1	0,2959	77	*granuloma*	1	0,2959
38	*posterioranterior*	1	0,2959	78	*present*	1	0,2959
39	*lordotic*	1	0,2959	79	*behind*	1	0,2959
40	*views*	1	0,2959	80	*heart*	1	0,2959
41	*pleura*	1	0,2959	81		-	-

de predizer z_{T-1} (ou seja, remover o ruído). Esse processo de remoção de ruído é repetido em um ciclo de $T-1$ vezes ou passos até chegar em z (imagem compacta). Após isso, o *Decoder* (D) descompacta a imagem e então a retorna do espaço latente para o espaço de pixel.

- Além disso, a abordagem permite fazer o condicionamento (*Conditioning*) arbitrário mapeando várias modalidades de entrada, como mapas semânticos (*Semantic Map*), texto (*Text*) ou imagem (as modalidades de entrada são apresentadas na parte superior do bloco direito da Figura 3.4, representadas por $\boldsymbol{y}$). Em outras palavras, os LDMs são, em princípio, capazes de modelar (aprender) distribuições condicionais da forma $p(z|y)$ (lê-se: dado y, aprender a distribuição ou representação latente z). Onde z é a representação latente da imagem, ε= *encoder*, x= imagem e z= ε(x). Para pré-processar a entrada y a partir de várias modalidades (como *prompts*), lado

Tabela 3.2: Os 100 termos mais frequentes no *dataset* de 30 imagens.

	Termos	Ocorrência	Porcentagem (%)		Termos	Ocorrência	Porcentagem (%)
1	*patients*	60	13,3630	51	*cavitation*	2	0,4454
2	*years*	32	7,1269	52	*apex*	2	0,4454
3	*sex*	30	6,6815	53	*smears*	2	0,4454
4	*age*	30	6,6815	54	*area*	2	0,4454
5	*normal*	16	3,5635	55	*visible*	2	0,4454
6	*f*	13	2,8953	56	*mcucxr01621*	2	0,4454
7	*lobe*	12	2,6726	57	*similar*	2	0,4454
8	*tuberculosis*	11	2,4499	58	*prior*	2	0,4454
9	*left*	11	2,4499	59	*views*	2	0,4454
10	*upper*	7	1,5590	60	*2m*	2	0,4454
11	*active*	7	1,5590	61	*slight*	2	0,4454
12	*positive*	6	1,3363	62	*improvement*	2	0,4454
13	*right*	6	1,3363	63	*isoniazid*	2	0,4454
14	*patient*	5	1,1136	64	*h*	2	0,4454
15	*ago*	5	1,1136	65	*rifampicin*	2	0,4454
16	*pleural*	5	1,1136	66	*r*	2	0,4454
17	*disease*	4	0,8909	67	*pyrazinamide*	2	0,4454
18	*chest*	4	0,8909	68	*z*	2	0,4454
19	*xray*	4	0,8909	69	*ethambutol*	2	0,4454
20	*effusion*	4	0,8909	70	*scars*	1	0,2227
21	*since*	4	0,8909	71	*unchanged*	1	0,2227
22	*large*	4	0,8909	72	*treated*	1	0,2227
23	*1m*	4	0,8909	73	*directly*	1	0,2227
24	*extensive*	3	0,6682	74	*observed*	1	0,2227
25	*cavitary*	3	0,6682	75	*involving*	1	0,2227
26	*culture*	3	0,6682	76	*radiation*	1	0,2227
27	*infiltrate*	3	0,6682	77	*year*	1	0,2227
28	*consistent*	3	0,6682	78	*contact*	1	0,2227
29	*6m*	3	0,6682	79	*case*	1	0,2227
30	*infiltrates*	3	0,6682	80	*bilateral*	1	0,2227
31	*cavity*	3	0,6682	81	*miliary*	1	0,2227
32	*inactive*	2	0,4454	82	*nodules*	1	0,2227
33	*previously*	2	0,4454	83	*diffusely*	1	0,2227
34	*therapy*	2	0,4454	84	*middle*	1	0,2227
35	*pansensitive*	2	0,4454	85	*present*	1	0,2227
36	*smear*	2	0,4454	86	*behind*	1	0,2227
37	*lingula*	2	0,4454	87	*heart*	1	0,2227
38	*areas*	2	0,4454	88	*stable*	1	0,2227
39	*lung*	2	0,4454	89	*prescription*	1	0,2227
40	*well*	2	0,4454	90	*mcucxr01131*	1	0,2227
41	*tuberculin*	2	0,4454	91	*rx*	1	0,2227
42	*skin*	2	0,4454	92	*psoas*	1	0,2227
43	*test*	2	0,4454	93	*abscess*	1	0,2227
44	*old*	2	0,4454	94	*abdominal*	1	0,2227
45	*calcified*	2	0,4454	95	*lymphadenopathy*	1	0,2227
46	*granuloma*	2	0,4454	96	*iodoamphetamine*	1	0,2227
47	*lower*	2	0,4454	97	*near*	1	0,2227
48	*findings*	2	0,4454	98	*hilum*	1	0,2227
49	*likely*	2	0,4454	99	*superior*	1	0,2227
50	*small*	2	0,4454	100	*segment*	1	0,2227

direito da Figura 3.4, é introduzido um codificador τ_θ (*encoder*) que projeta a entrada y para uma representação intermediária $\tau_\theta(y)$, que é então mapeada para as camadas intermediárias da Rede Neural Convolucional UNet por meio de uma camada de atenção cruzada.

Neste trabalho, a tarefa principal consiste em realizar o *fine-tuning* do *Latent Diffusion Foundation Model* (LDFM) visando gerar imagens sintéticas de raios-x do tórax de alta resolução. Para isso, recorremos ao uso da interface gráfica kohya-ss GUI, desenvolvida em Python. Esta interface é compatível tanto com a interface do usuário do *Stable Diffusion Model* (LoRA [43]), quanto com a interface gráfica utilizada para a geração de

Tabela 3.3: Os 100 termos mais frequentes no *dataset* de 100 imagens.

	Termos	Ocorrência	Porcentagem (%)		Termos	Ocorrência	Porcentagem (%)
1	*patients*	200	11,5607	51	*scarring*	4	0,2312
2	*chest*	111	6,4162	52	*pulmonary*	4	0,2312
3	*xray*	109	6,3006	53	*improving*	3	0,1734
4	*years*	109	6,3006	54	*cavitation*	3	0,1734
5	*human*	104	6,0116	55	*views*	3	0,1734
6	*sex*	100	5,7803	56	*nodular*	3	0,1734
7	*age*	100	5,7803	57	*acid*	3	0,1734
8	*normal*	51	2,9480	58	*bilateral*	3	0,1734
9	*healthy*	50	2,8902	59	*granuloma*	3	0,1734
10	*f*	49	2,8324	60	*apical*	3	0,1734
11	*lobe*	42	2,4277	61	*rt*	3	0,1734
12	*tuberculosis*	35	2,0231	62	*apex*	3	0,1734
13	*upper*	32	1,8497	63	*pt*	3	0,1734
14	*right*	32	1,8497	64	*prior*	3	0,1734
15	*left*	27	1,5607	65	*rifampicin*	3	0,1734
16	*active*	17	0,9827	66	*r*	3	0,1734
17	*infiltrate*	15	0,8671	67	*pyrazinamide*	3	0,1734
18	*inactive*	14	0,8092	68	*z*	3	0,1734
19	*pleural*	13	0,7514	69	*ethambutol*	3	0,1734
20	*ago*	12	0,6936	70	*observed*	3	0,1734
21	*scars*	12	0,6936	71	*therapy*	3	0,1734
22	*disease*	11	0,6358	72	*culture*	3	0,1734
23	*consistent*	10	0,5780	73	*pansensitive*	3	0,1734
24	*cavitary*	10	0,5780	74	*similar*	3	0,1734
25	*lung*	9	0,5202	75	*since*	3	0,1734
26	*treatment*	9	0,5202	76	*volume*	3	0,1734
27	*effusion*	9	0,5202	77	*treated*	3	0,1734
28	*large*	8	0,4624	78	*stable*	3	0,1734
29	*positive*	8	0,4624	79	*ray*	2	0,1156
30	*likely*	7	0,4046	80	*shows*	2	0,1156
31	*lower*	7	0,4046	81	*view*	2	0,1156
32	*infiltrates*	6	0,3468	82	*lordotic*	2	0,1156
33	*findings*	6	0,3468	83	*appearance*	2	0,1156
34	*6m*	6	0,3468	84	*radiotherapy*	2	0,1156
35	*well*	5	0,2890	85	*also*	2	0,1156
36	*isoniazid*	5	0,2890	86	*bilaterally*	2	0,1156
37	*history*	5	0,2890	87	*smears*	2	0,1156
38	*x*	4	0,2312	88	*mycobacterium*	2	0,1156
39	*middle*	4	0,2312	89	*small*	2	0,1156
40	*extensive*	4	0,2312	90	*areas*	2	0,1156
41	*cavity*	4	0,2312	91	*tuberculin*	2	0,1156
42	*old*	4	0,2312	92	*skin*	2	0,1156
43	*calcified*	4	0,2312	93	*test*	2	0,1156
44	*patient*	4	0,2312	94	*miliary*	2	0,1156
45	*h*	4	0,2312	95	*nodules*	2	0,1156
46	*smear*	4	0,2312	96	*heart*	2	0,1156
47	*improvement*	4	0,2312	97	*area*	2	0,1156
48	*loss*	4	0,2312	98	*visible*	2	0,1156
49	*prescription*	4	0,2312	99	*lesser*	2	0,1156
50	*1m*	4	0,2312	100	*year*	2	0,1156

imagens, AUTOMATIC1111 [5]. Esta interface gráfica é especialmente útil para a geração de imagens, pois permite realizar *fine-tuning* do LDFM.

Um dos recursos oferecidos pela kohya-ss GUI é o LoRA (*Low-Rank Adaptation*), proposto por Hu et al. [43]. Este recurso possibilita a redução do número de parâmetros treináveis em até 10.000 vezes e do requisito de memória da GPU em até 3 vezes, quando comparado ao GPT-3 175B (GPT-3 com 175 bilhões de parâmetros) ajustado com o otimizador Adam8bit.

Existem diversas abordagens para a geração de imagens personalizadas. O *fine-tuning* do LDFM utilizando abordagens como DreamBooth [91] e LoRA [43] são exemplos de métodos possíveis, representativos e amplamente usados [35]. Em termos de desem-

Tabela 3.4: Ocorrência dos termos utilizados nos *prompts* de geração de imagens no conjunto de dados de 30 imagens. O conjunto de dados de 30 imagens é um dos conjuntos utilizados para o treinamento dos modelos na primeira fase.

Termos	**Ocorrência dos termos no *dataset* de 30 imagens**	**Porcentagem de ocorrência dos termos no *dataset* de 30 imagens(%)**
Pleural	5	1,1136
Effusion	4	0,8909
Pleural effusion	0	0,0000
Infiltrate	3	0,6682
Infiltrates	3	0,6682
Nodular	0	00000
Nodules	1	0,2227
Nodular infiltrate	0	0,0000
Cavitation	2	0,4454
Tuberculosis	11	2,4499

Tabela 3.5: Ocorrência dos termos utilizados nos *prompts* de geração de imagens nos conjuntos de dados de 20 e 100 imagens. Os conjuntos de dados de 20 e 100 imagens são os conjuntos utilizados para o treinamento dos modelos 1 e 2 na segunda fase.

Termos	**Ocorrência dos termos nos *datasets* de 20 e 100 imagens**		**Porcentagem de ocorrência dos termos nos *datasets* (%)**	
	***dataset* de 20 imagens**	***dataset* de 100 imagens**	***dataset* de 20 imagens**	***dataset* de 100 imagens**
Pleural	7	13	2,0710	0,7514
Effusion	7	9	2,0710	0,5202
Pleural effusion	0	0	0,0000	0,0000
Infiltrate	7	15	2,0710	0,8671
Infiltrates	2	6	0,5917	0,3468
Nodular	2	3	0,5917	0,1734
Nodules	1	2	0,2959	0,1156
Nodular infiltrate	0	0	0,0000	0,0000
Cavitation	2	3	0,5917	0,1734
Tuberculosis	10	35	2,9586	2,0231

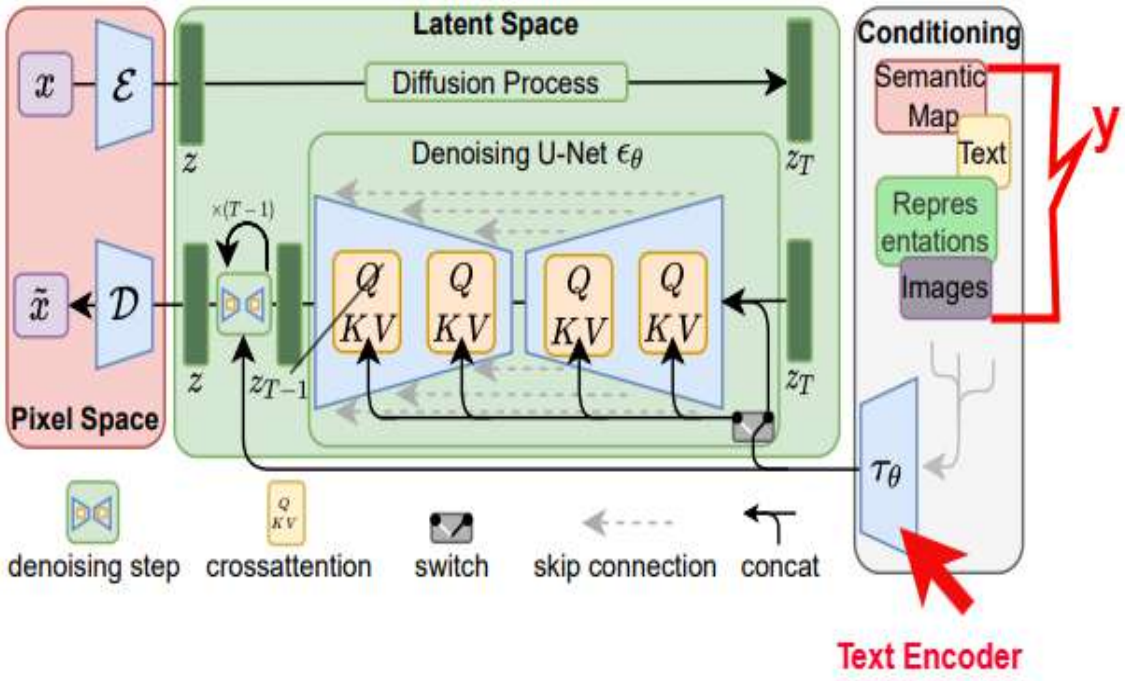


Figura 3.4: Estrutura de *Latent Diffusion Model* [89].

penho, o LoRA se mostra mais eficiente para treinar os modelos e compartilha-los entre os usuários quando comparado ao DreamBooth, que armazena todos os parâmetros do modelo após o treinamento [43, 91].

3.3.1 Interfaces de Treinamento e Geração de Imagens

Para agilizar o processo de treinamento e realizar o *fine-tuning* no *Stable Diffusion Foundation Model*, além de aprimorar a geração de imagens, foram empregadas diversas ferramentas para auxiliar nessas tarefas.

3.3.2 Interface para Treinamento do Modelo

Para tornar o processo de treinamento mais acessível, foi utilizada a interface gráfica de usuário chamada Kohya-ss (Figura 3.5). Isso permitiu configurar os parâmetros de treinamento e executar comandos essenciais para o treinamento do modelo. A seguir, especificam-se detalhes dos métodos utilizados:

kohya-ss GUI

A interface kohya-ss GUI é uma ferramenta que facilita a implementação de *scripts* de treinamento do *Stable Diffusion Model* [9]. Ela oferece a opção de uso do LoRA (*Low Rank Adaptation*) como interface para o treinamento do modelo. Embora a kohya-ss GUI permita uma variedade de métodos de treinamento, como o Dreambooth (indicado pela seta azul no canto superior esquerdo da Figura 3.5), o treinamento usando LoRA (indicado pela seta verde no canto superior esquerdo da Figura 3.7) é o mais adequado para este trabalho.

O LoRA se destaca por tornar o treinamento mais eficiente, reduzindo as exigências de recursos de hardware em até três vezes ao usar otimizadores adaptativos, como o otimizador Adam8bit. Isso ocorre porque, com o LoRA, não precisamos calcular os gradientes ou manter os estados do otimizador para a maioria dos parâmetros. Em vez disso, otimizamos apenas as matrizes de posto baixo injetadas, que são muito menores. Isso reduz significativamente o número de parâmetros treináveis e os requisitos de memória da GPU, uma vez que os gradientes não precisam ser calculados para a maioria dos pesos do modelo [43].

Além disso, o LoRA se destaca no quesito compatibilidade, pois pode ser aplicado às camadas de atenção cruzada que relacionam as representações da imagem com os

prompts que as descrevem. Isso o torna uma ferramenta versátil para realizar *fine-tuning* de vários modelos [22].

Quanto ao tamanho do modelo, os pesos treinados são muito menores com o LoRA, o que o torna um modelo mais fácil de gerenciar e compartilhar [22].

LoRA

LoRA é um método utilizado para fazer *fine-tuning* do *Stable Diffusion Models* [43], fornecendo imagens de treinamento e legendas de imagens associadas, semelhante ao funcionamento do Dreambooth. Ele permite que o usuário direcione (influencie ou enviese) *Stable Diffusion Model* em direção às imagens fornecidas. Sendo assim, este software será usado para adaptar *Stable Diffusion Model* às imagens de raios-x de tórax [43].

O LoRA também possibilita congelar os pesos do modelo pré-treinado e incorporar matrizes de decomposição de classificação treináveis em cada camada da arquitetura do Transformer. Isso resulta em uma redução significativa no número de parâmetros treináveis, tornando o modelo mais eficiente para tarefas subsequentes [43].

Outro aspecto relevante é a vantagem do LoRA no quesito eficiência. Os benefícios do LoRA em relação ao Dreambooth são os baixos requisitos em termos de tempo e recurso computacional. Treinar um modelo com LoRA pode levar algumas horas, em comparação com dias quando são utilizados outros métodos como o Dreambooth. Quanto a exigência do poder computacional, LoRA tem vantagem sobre o Dreambooth por exigir menos recursos computacionais.

TensorBoard

A interface Kohya-ss também permite usar uma importante ferramenta de visualização de gráficos denominada TensorBoard (seta vermelha, Figura 3.5). O TensorBoard permite exibir gráficos do modelo, plotar valores escalares e seriais à medida que o treinamento progride. Na Figura 3.6, é apresentado um exemplo de visualização das métricas de *loss* e a taxa de aprendizado (*Learning rate* - *lr*). Da esquerda para a direita e de cima para baixo, temos:

- Gráfico da perda média (*loss/average*): nele, é exibida a média dos valores da função de perda ao longo de um certo número de iterações ou épocas durante o processo de treinamento do modelo. Isso nos dá uma ideia geral do desempenho do modelo ao longo do tempo. Uma perda média decrescente, como ocorre no gráfico, indica que o modelo está aprendendo e melhorando.
- Gráfico da perda atual (*loss/current*): nele, é exibido o valor da função de perda para a iteração ou época mais recente durante o processo de treinamento do modelo. Isso

nos dá *feedback* imediato sobre o desempenho do modelo no último lote de dados em que foi treinado. Uma perda atual decrescente, como ocorre também no gráfico, indica que o modelo está aprendendo e melhorando.

- Gráfico de perda por época (*loss/epoch*): nele, é exibida a perda média em uma passagem completa por todo o conjunto de dados de treinamento. Uma época trata-se de uma passagem completa por todo o conjunto de dados de treinamento. Normalmente, a perda por época é calculada como a perda média em todos os lotes do conjunto de dados. Essa métrica é particularmente útil para monitorar o processo de treinamento, porque fornece uma visão mais abrangente do desempenho do modelo em comparação à perda atual, que reflete apenas o desempenho do lote mais recente. Neste caso também, uma perda decrescente, como ocorre no gráfico, indica que o modelo está aprendendo e melhorando.

- Gráfico da taxa de aprendizagem do codificador de texto (*lr/textencoder*): nele, é exibida taxa de aprendizagem usada ao treinar a parte do codificador de texto do modelo. O codificador de texto é um componente do modelo que transforma dados de texto em uma representação numérica (onde a representação numérica pode ser um vetor ou uma sequência de vetores) que o modelo pode compreender e processar. A taxa de aprendizagem é um hiperparâmetro que determina o quanto os pesos do modelo são atualizados durante o treinamento, enquanto se move em direção ao mínimo de uma função de perda. Uma taxa de aprendizagem adequada ajuda o modelo a aprender de forma eficaz com os dados, ou seja, ele decide o quão rápido ou lento o modelo aprende.

- Gráfico da taxa de aprendizagem da Unet (*lr/unet*): nele, é exibida a taxa de aprendizagem usada ao treinar a Unet do modelo. A Unet é uma de rede neural convolucional frequentemente usada para tarefas de segmentação de imagens. Definir a taxa de aprendizagem da Unet, significa determinar o quanto os pesos dessa rede são atualizados durante o treinamento do modelo.

3.3.3 Procedimento de Treinamento do Modelo

Esta seção apresenta os procedimentos utilizados para treinar o modelo utilizando LoRA. O treinamento foi realizado utilizando diferentes otimizadores tais como: **AdamW8bit** (utilizando configuração padrão), **AdamW8bit** (alterando configurações), **Adafactor**, **DAdaptSGD** e **Prodigy**, conforme detalhados na Tabela 3.6. O treinamento envolve duas fases:

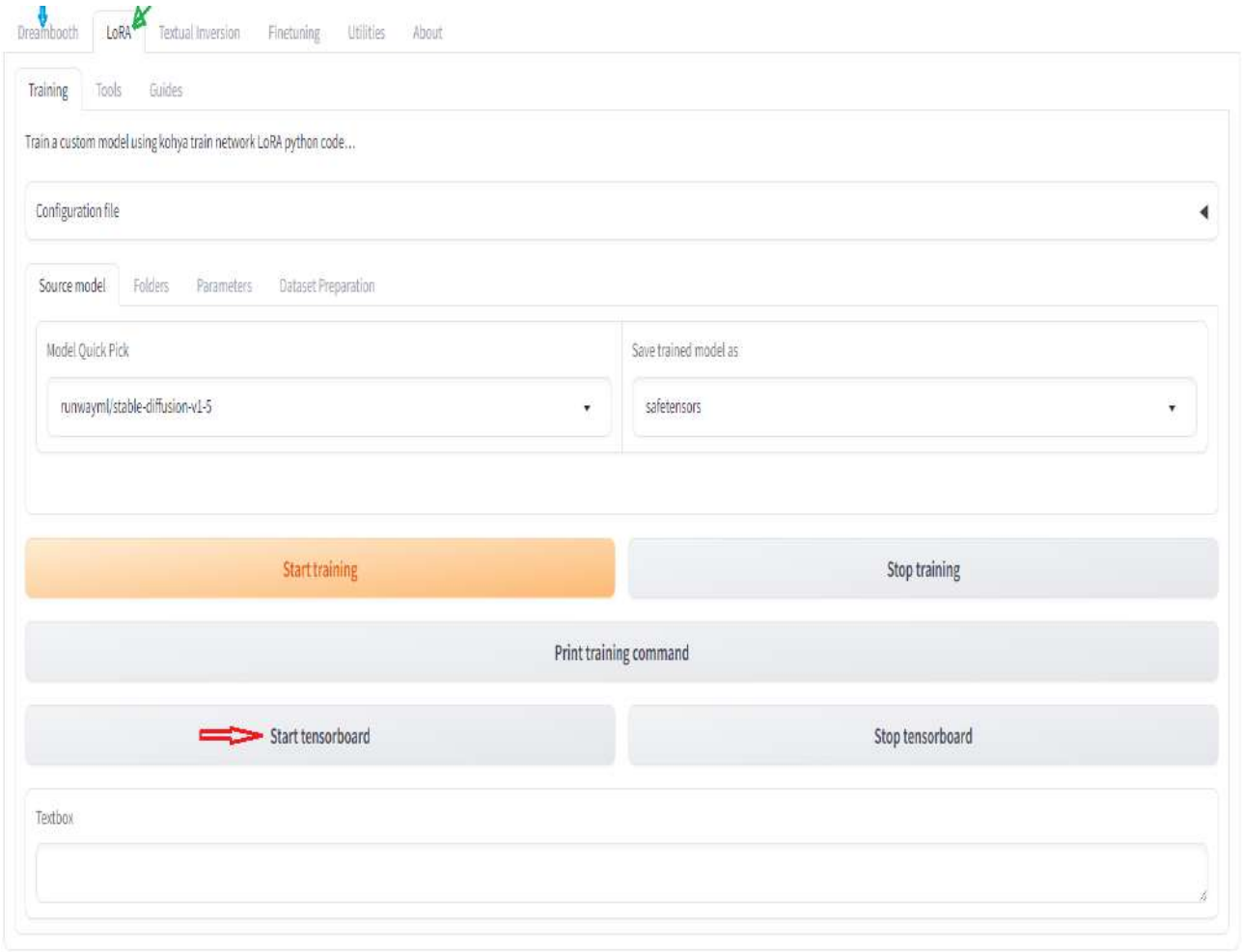


Figura 3.5: Interface kohya-ss utilizado para o treinamento dos modelos [9].

- Primeira fase: o treinamento foi realizado utilizando quatro otimizadores diferentes (Tabela 3.6). O conjunto de dados utilizado para treinar os modelos conta com 30 imagens, 15 imagens saudáveis e 15 imagens não saudáveis, ambas com os respectivos relatórios clínicos. Nesta fase, as imagens geradas foram avaliadas por um único especialista.

- Segunda fase: na segunda fase do treinamento, utilizamos exclusivamente o otimizador Adam8bit, que se destacou como o melhor entre os modelos treinados na primeira fase. As configurações do Modelo 1 e Modelo 2, apresentadas na Tabela 3.6, foram empregadas nesta etapa. Nesta fase, trabalhamos com conjuntos de dados compostos por 5, 10, 20, 40, 60, 80 e 100 imagens. Em cada conjunto, mantivemos uma proporção equilibrada, com 50% de imagens saudáveis e 50% de imagens não saudáveis, todas acompanhadas de seus respectivos relatórios clínicos. Contudo, devido à escassez de especialistas para avaliar a grande quantidade de imagens geradas, optamos por selecionar apenas as imagens geradas pelos modelos treinados com os conjuntos de dados de 20 e 100 imagens. Essa estratégia foi adotada

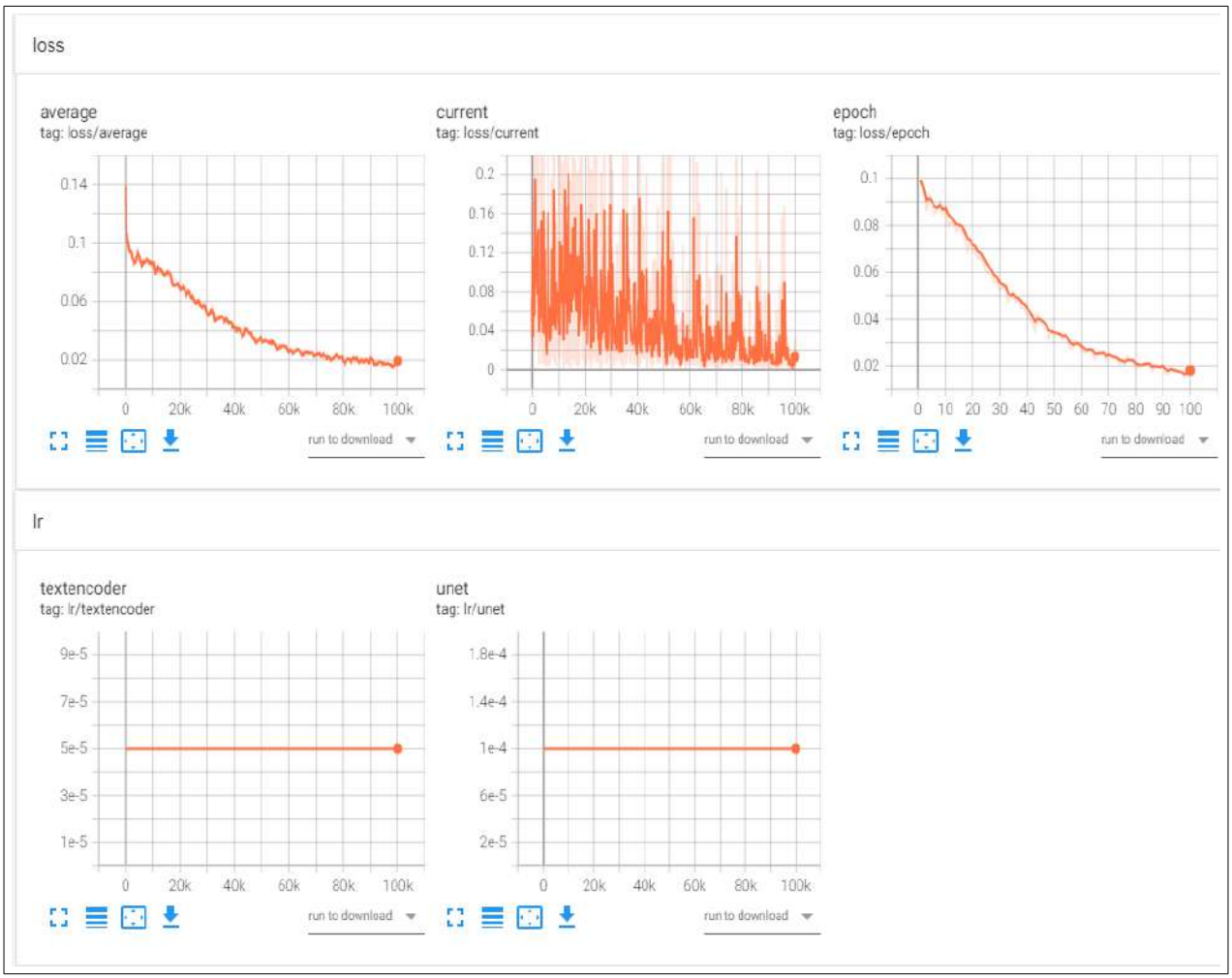


Figura 3.6: Visualização de métricas de *loss* e taxa de aprendizado (lr) à medida que o treinamento progride, utilizando a ferramenta TensorBoard [9].

para tornar o processo de avaliação mais viável. É importante frisar que as imagens geradas nesta fase forma avaliadas por duas médicas.

Em ambas as fases, o treinamento foi realizado por 100 épocas e para cada modelo treinado, os submodelos foram sendo gerados a cada 20 épocas. Adotamos essa estratégia para nos possibilitar testar o número de épocas suficientes para treinar um modelo capaz de gerar imagens de qualidade aceitável. Isso é útil, uma vez que permite ter noção da quantidade de imagens e diversas suficientes para treinar um modelo capaz de produzir melhores resultados.

Os otimizadores apresentados na Tabela 3.6 seguem abordados com mais detalhes nos itens elencados abaixo:

Adam

O otimizador Adam é um método de otimização estocástica adequado para problemas de aprendizado de máquina em grande escala. Ele adapta a taxa de aprendizagem

Tabela 3.6: Parâmetros utilizados na configuração do *software* LoRA para o treinamento dos modelos. O otimizador adam8bit, que é destacado em azul, foi empregado em duas situações distintas. Na primeira, mantivemos as configurações padrão do otimizador adam8bit e procedemos com o treinamento do modelo. Este modelo, treinado com as configurações padrão do otimizador adam8bit, é denominado Modelo 01. Na segunda situação, fizemos alterações nas configurações do otimizador adam8bit e treinamos um novo modelo. Este modelo, treinado com as configurações modificadas do otimizador adam8bit, é referido como Modelo 02.

Modelos	Modelos 1	Modelos 2	Modelos 3	Modelos 4	Modelos 5
Otimizador	AdamW8bit	AdamW8bit	Adafactor	DAdaptSGD	Prodigy
Clip Skip	2	2	2	2	2
Época	100	100	100	100	100
LR	1.10^{-4}	1.10^{-4}	1.10^{-4}	1.10^{-4}	1.10^{-4}
Max resolution	152x512	152x512	152x512	152x512	152x512
LR Scheduler	constant	constant	constant	constant	constant
Pretrained Model	stable-diffusion-v1-5	stable-diffusion-v1-5	stable-diffusion-v1-5	stable-diffusion-v1-5	stable-diffusion-v1-5
Train batch size	2	2	2	2	2
Text encoder lr	5.10^{-5}	5.10^{-5}	5.10^{-5}	1.10^{-5}	1.10^{-5}
Unet lr	1.10^{-4}	1.10^{-4}	1.10^{-4}	1.10^{-5}	1.10^{-5}
vae batch size	0	32	32	32	32
noise offset type	Original	Multires	Multires	Multires	Multires
multires noise discount	0	0.1	0.1	0.1	0.1
multires noise iteration	0	6	6	6	6

para cada peso do modelo individualmente e calcula taxas de aprendizagem adaptativas para diferentes parâmetros. É um otimizador invariante ao redimensionamento diagonal dos gradientes, ou seja, ele se comporta da mesma maneira mesmo que os gradientes sejam multiplicados por uma matriz diagonal com apenas fatores positivos [56].

O otimizador **Adam8bit** é uma versão do otimizador Adam. Ele é um otimizador de 8 bits, o que lhe proporciona maior rapidez no treinamento do modelo (4x mais rápido que o Adam [30]) e menor uso de memória (podendo reduzir de 33 a 75 porcento do consumo total de memória durante o treinamento [26]), mantendo o desempenho de otimizadores de 32 bits [26].

Adafactor

Nos métodos de otimização estocástica como RMSProp, Adam e Adadelta, as atualizações de parâmetros são escalonadas pelas raízes quadradas inversas de médias móveis exponenciais de gradientes quadrados. Porém, manter esses estimadores de segundo momento por parâmetro requer memória igual ao número de parâmetros, o que pode ser impeditivo para modelos grandes. O Adafactor, por sua vez, resolve essa demanda mantendo apenas as somas por linha e por coluna dessas médias móveis e estimando os segundos momentos por parâmetro com base nessas somas. Com isso, o

Adafactor reduz significativamente o requisito de memória, mantendo os benefícios da adaptabilidade, tornando viável o treinamento de modelos maiores [92].

DAdaptSGD

D-Adaption ou DAdaption é uma abordagem para definir automaticamente a taxa de aprendizagem que atinge assintoticamente a taxa ideal de convergência para minimizar funções convexas de Lipschitz, sem back-tracking ou pesquisas de linha, e sem valor de função adicional ou avaliações de gradiente por etapa. Por outro lado, DAdaptSGD é uma variante do DAdaption voltada a definir automaticamente a taxa de aprendizagem no algoritmo de otimização *Stochastic Gradient Descent* (SGD) [25].

Prodigy

O Prodigy é um otimizador que aborda o problema de estimativa da taxa de aprendizagem em métodos adaptativos, como Adagrad e Adam. Ele é capaz de estimar a distância até a solução, necessária para definir a taxa de aprendizagem de maneira ideal. É uma técnica resultante de modificações do método D-Adaptation para *learning-rate-free learning* (um conceito de aprendizado de máquina em que a taxa de aprendizagem é definida automaticamente, e não manualmente) [72].

Todos os cinco modelos apresentados na Tabela 3.6 foram treinados utilizando o *Foundation Model* stable-diffusion-v1-5 como modelo base. O *Foundation Model* stable-diffusion-v1-5 é um modelo de difusão de texto para imagem capaz de gerar imagens fotorrealistas a partir de entrada de texto. Ele foi desenvolvido por Rombach et al. [89], e foi treinado no conjunto de dados denominado **laion-aesthetics v2 5+** [31]. O **laion-aesthetics v2 5+** é um subconjunto do conjunto de dados **LAION 5B** [31], com os subconjuntos de amostras com legenda em inglês como apresentado abaixo:

- 1,2 bilhões de pares de imagem-texto com pontuações estéticas de 4,5 ou superiores;
- 939 milhões de pares de imagem-texto com pontuações estéticas de 4,75 ou superiores;
- 600 milhões de pares de imagem-texto com pontuações estéticas de 5 ou superiores;
- 12 milhões de pares de imagem-texto com pontuações estéticas de 6 ou superiores;
- 3 milhões de pares de imagem-texto com pontuações estéticas de 6,25 ou superiores;
- 625 mil pares de imagem-texto com pontuações estéticas de 6,5 ou superiores.

3.3.4 Interface para a Geração de Imagens

Finalmente, para a geração de imagens, foi empregada a interface AUTOMATIC1111 [5], que é uma WebUI desenvolvida para trabalhar com o modelo *Stable Diffusion*. Essa interface permite inserir diversas configurações e gerar imagens desejadas, consolidando, assim, o ciclo de treinamento e geração de imagens proposto neste trabalho.

AUTOMATIC1111

AUTOMATIC1111 é uma WebUI que visa disponibilizar parâmetros aos usuários em uma interface web abrangente. Na Figura 3.7 é possível ver uma interface composta por caixas de texto, controles deslizantes e botões com diferentes opções. Para explorar o efeito de um parâmetro, o usuário precisa alterar manualmente o parâmetro e gerar um novo conjunto de imagens com a nova configuração. A interface também permite que o usuário altere as configurações de parâmetros para gerar uma grande variedade de imagens de uma só vez.

O processo de geração de imagem utilizando a interface AUTOMATIC1111 [5] ocorre da seguinte forma: primeiro é escolhido o modelo que se pretende utilizar para gerar a imagem e em seguida o *prompt* descrevendo o tipo de imagem que se pretende gerar (Figura 3.7). A interface AUTOMATIC1111 permite também que o usuário adicione configurações adicionais à interface de geração de imagens.

A Figura 3.8 ilustra o fluxo simplificado de geração de imagem tendo como base o modelo proposto por Rombach et al. [89]. Basicamente, introduz-se o *prompt* de entrada com as características da imagem que se pretende gerar. No exemplo da figura em questão, utilizamos o *prompt* de geração de imagem saudável "*Healthy chest x-ray*". O codificador de texto (***Text Encoder***) por sua vez, faz o papel de converter (codificar) a entrada textual em um vetor latente (em uma representação vetorial). O restante do processo é elencado nos itens abaixo:

- ***Conditioning***: é o processo no qual é fornecido informações adicionais (como texto ou imagens) para orientar o modelo de difusão latente para gerar imagens que sejam relevantes e consistentes com as informações fornecidas.
- Ruído aleatório (***Random Noise***): Para produzir distribuições sobre imagens possíveis, o ruído aleatório é adicionado ao espaço latente. Geralmente, esse ruído é do tipo gaussiano e serve basicamente para introduzir a estocasticidade no processo de geração da imagem.

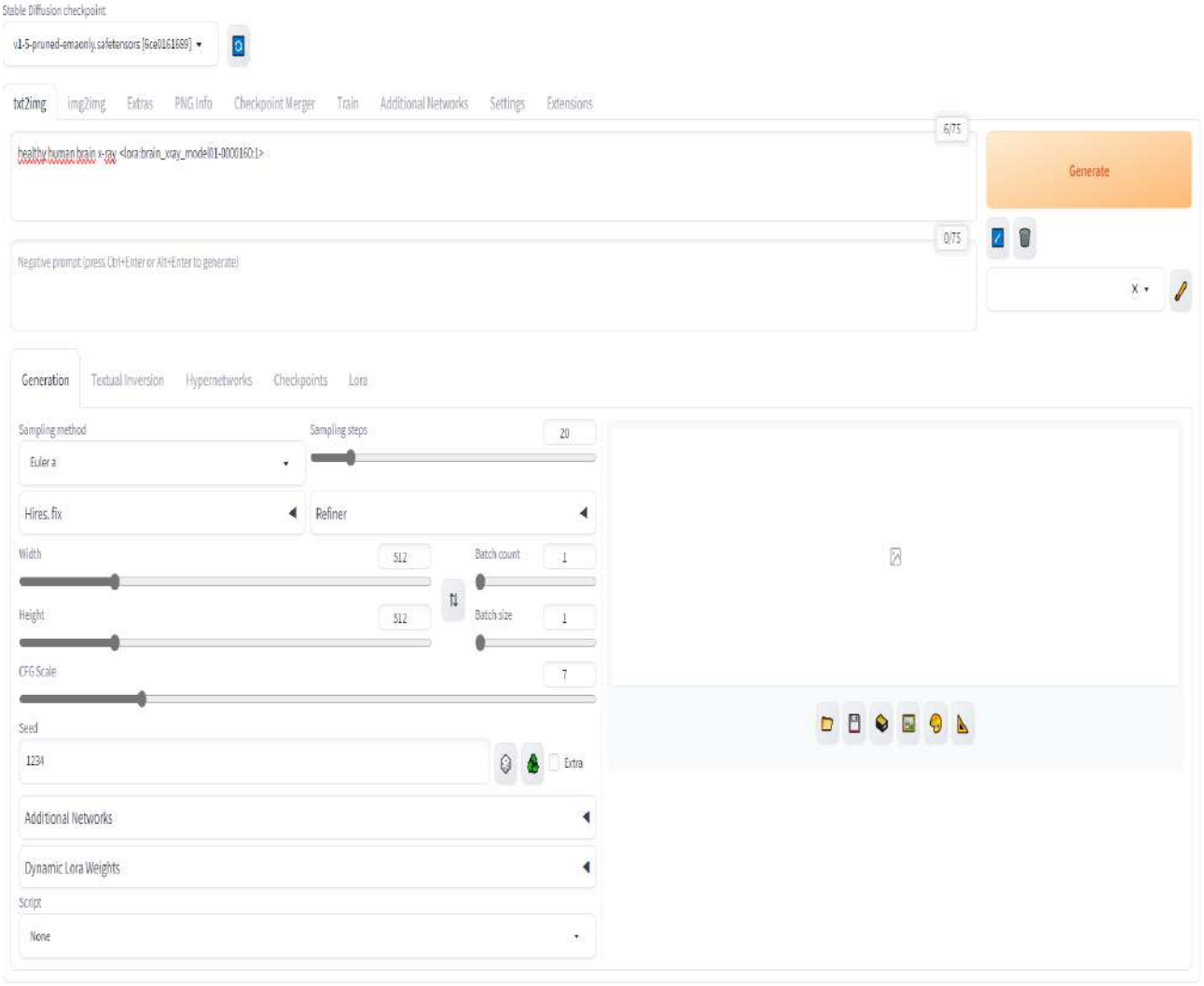


Figura 3.7: Interface do *software* AUTOMATIC1111 utilizado para a geração de imagens [5].

- Na sequência, ***Denoising Unet*** é usada para aprender as etapas de adição e remoção de ruído. A ***Denoising Unet*** recebe dados como ruído, informações de intervalo de tempo e algum sinal de condicionamento (como o vetor de texto) e prevê resíduos de ruído que podem ser aplicados para eliminar ruído da imagem.

- **Decodificador**: o decodificador por sua vez recebe uma representação vetorial da imagem e produz imagem. Ou seja, o decodificador é responsável por gerar imagens de alta resolução a partir do espaço latente.

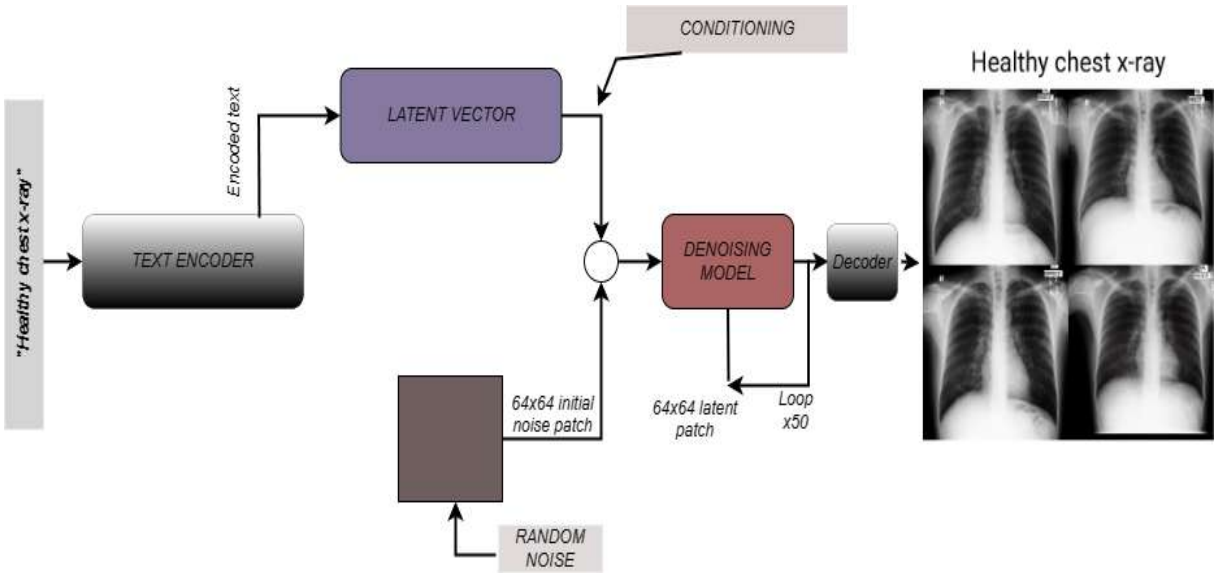


Figura 3.8: Fluxo de geração de imagens. Nesta figura, apresentamos um exemplo que utilizamos o *prompt* de geração de imagem saudável "*Healthy chest x-ray*" (lado esquerdo da figura). Após o processo de codificação do *prompt* (*TEXT ENCODER*), o condicionamento (*CONDITIONING*), o processo de remoção de ruído (*DENOISING*) e o processo de decodificação (*Decoding*), é gerado o conjunto formado por quatro imagens apresentados no lado direito da figura.

4. RESULTADOS

Este capítulo é dedicado à apresentação dos resultados alcançados no nosso estudo. Inclui discussões sobre os resultados que obtivemos e as avaliações realizadas por profissionais de saúde. O capítulo está organizado em quatro seções. Na Seção 4.1, apresentamos a primeira fase de avaliação das imagens geradas. Em seguida, na Seção 4.2, discutimos a segunda fase de avaliação dessas imagens. Avançamos para a Seção 4.3, onde é realizada uma discussão geral sobre os resultados. Por fim, na Seção 4.4, apresentamos os requisitos computacionais necessários para o estudo.

4.1 Primeira Fase de Avaliação das Imagens Geradas

Na primeira fase, as imagens foram geradas utilizando seis modelos diferentes: o modelo pré-treinado (denominado *Foundation Model*) e os cinco modelos treinados utilizando os quatro tipos de otimizadores apresentados na Tabela 3.6 (Adam8bit, uma versão com as configurações *default* e outra versão alterando configurações, Adafactor, DAdaptSGD e Prodigy). Para cada modelo treinado, foi utilizado um otimizador específico, com exceção dos Modelos 1 e 2 treinados utilizando o mesmo otimizador (Adam8bit), porém mantendo as configurações *default* no treinamento do Modelo 1 e alterando as configurações no caso do treinamento do Modelo 2, conforme as configurações apresentadas na Tabela 3.6 (colunas Modelo 1 e Modelo 2). Para a geração das imagens enviadas para a avaliação, foram utilizados os *prompts* apresentados na Tabela 4.1. O *prompt* "*healthy or normal human chest x-ray*", foi utilizado para gerar imagens de raios-x de tórax saudáveis e o *prompt* "*Human chest x-ray with tuberculosis. Bilateral miliary nodules with Right Middle Lobe infiltrate. Right pleural effusion*", para gerar imagens de raios-x de tórax não saudáveis.

Nas Figuras 4.1 e 4.2, apresentamos as imagens geradas por cada modelo. Todos os modelos utilizados para gerá-las foram treinados por cem épocas. Os conjuntos de imagens apresentados na Figura 4.1 foram gerados utilizando o *prompt* de geração de imagens de raios-x saudáveis "*healthy or normal human chest x-ray*". Por outro lado, os conjuntos de imagens apresentados na Figura 4.2 foram gerados utilizando o *prompt* de geração de imagens de raios-x não saudáveis "*Human chest x-ray with tuberculosis. Bi-lateral miliary nodules with Right Middle Lobe infiltrate. Right pleural effusion*".

As imagens geradas nessa primeira fase foram avaliadas por um médico utilizando as escalas e os fatores apresentados na Tabela 4.2. De acordo com avaliações do médico (Figura 4.3), as imagens geradas utilizando os modelos treinados com o otimizador Adam8bit são as mais realistas. Ou seja, os Modelos 1 e 2, cujas configurações

são apresentadas na Tabela 3.6, foram capazes de, dada uma entrada de texto (*prompt*), gerar imagens realistas.

Tabela 4.1: *Prompts* utilizados para gerar imagens de raios-x do tórax saudáveis e não saudáveis.

Prompt para geração de imagens saudáveis	*Prompts* para geração de imagens não saudáveis
"*healthy or normal human chest x-ray*"	"*Human chest x-ray with tuberculosis*. *Bilateral miliary nodules with Right Middle Lobe infiltrate*. *Right pleural effusion*"

Tabela 4.2: Escalas e fatores de avaliação das imagens geradas na primeira fase dos testes.

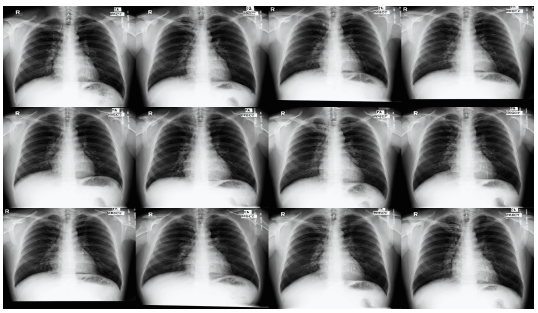
Escala	Fatores
1	Totalmente NÃO realista
2	POUCO realista
3	RAZOAVELMENTE realista
4	BOM Realismo
5	MUITO realista

4.1.1 Análise dos Gráficos de Avaliação de Desempenho dos Modelos

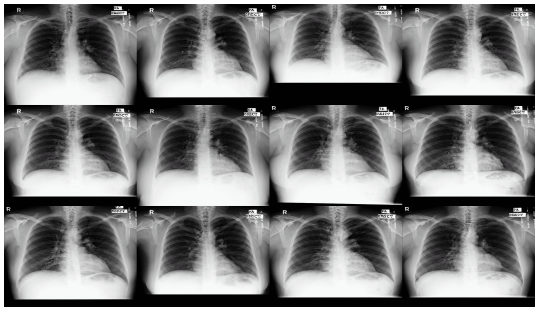
As imagens geradas (Figuras 4.1 e 4.2) pelos modelos apresentados na Tabela 3.6, foram avaliadas por um médico e o resultado das avaliações segue apresentado no gráfico da Figura 4.3 (barras da cor verde: representam avaliações dos conjuntos de imagens saudáveis; barras da cor marrom: representam avaliações dos conjuntos de imagens não saudáveis). Esses gráficos representam avaliações de desempenho dos modelos na geração de imagens de raios-x do tórax.

O **Conjunto 1**, do gráfico apresentado na Figura 4.3 (barra da cor verde), representa a avaliação de desempenho do modelo *Foundation Model stable-diffusion-v1-5* na geração de imagens de raios-x saudáveis representadas como **Conjunto 1** na Figura 4.1a. Para esse conjunto de imagens, o avaliador avaliou-o como sendo "Totalmente NÃO Realista". Portanto, podemos afirmar, segundo a avaliação do médico, que o modelo *Foun-*

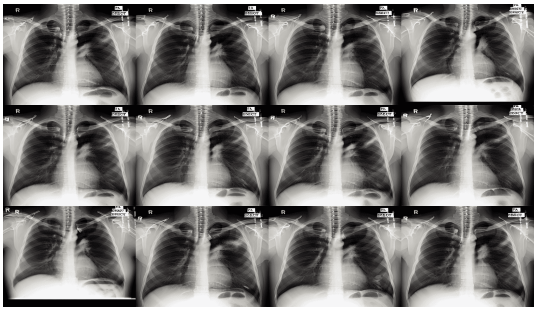
(a) **Conjunto 1**: Conjunto gerado pelo modelo pré-treinado *Foundation Model* stable-diffusion-v1-5 (**F.M**). O modelo foi treinado no conjunto de dados ***laion-aesthetics v2 5+*** [31].



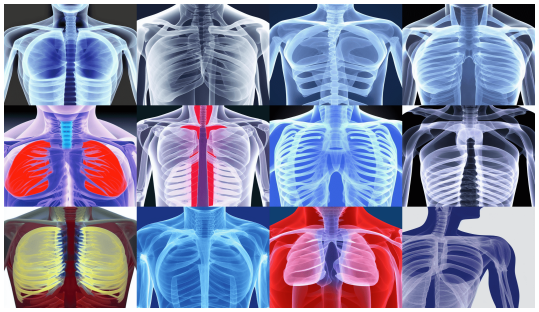
(b) **Conjunto 2**: Conjunto gerado pelo Modelo 01 (**M.1**). O modelo foi treinado utilizando o otimizador **Adam8bit**, utilizando as configurações do Modelo 01 apresentadas na Tabela 3.6.



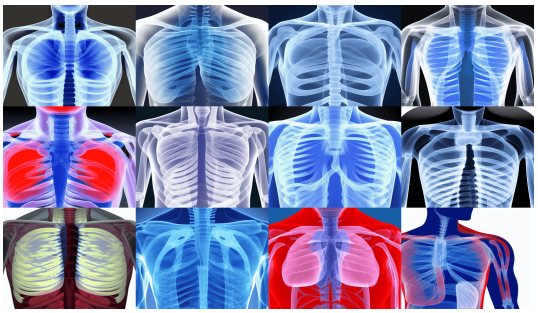
(c) **Conjunto 3**: Conjunto gerado pelo Modelo 02 (**M.2**). O modelo foi treinado utilizando o otimizador **Adam8bit**, utilizando as configurações do Modelo 02 apresentadas na Tabela 3.6.



(d) **Conjunto 4**: Conjunto gerado pelo Modelo 03 (**M.3**). O modelo foi treinado utilizando o otimizador **Adafactor**, utilizando as configurações do Modelo 03 apresentadas na Tabela 3.6.



(e) **Conjunto 5**: Conjunto gerado pelo Modelo 04 (**M.4**). O modelo foi treinado utilizando o otimizador **DAdaptSGD**, utilizando as configurações do Modelo 04 apresentadas na Tabela 3.6.

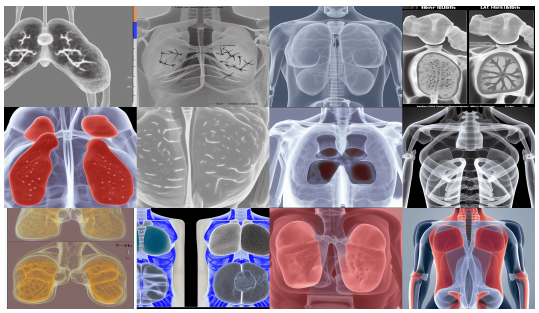


(f) **Conjunto 6**: Conjunto gerado pelo Modelo 05 (**M.5**). O modelo foi treinado utilizando o otimizador **Prodigy**, utilizando as configurações do Modelo 05 apresentadas na Tabela 3.6.

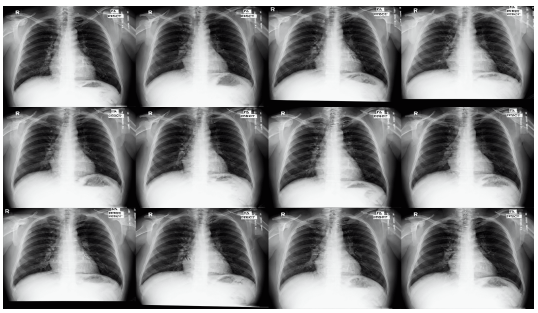
Figura 4.1: Imagens de raios-x do tórax geradas pelos modelos apresentados na Tabela 3.6 no processo de geração de imagens de raios-x saudáveis. Todas foram geradas pelos modelos treinados por 20 épocas em um conjunto de dados de 30 imagens de raios-x do tórax. Todas as imagens de raios-x foram geradas utilizando o *prompt* "*healthy or normal human chest x-ray*".

dation Model stable-diffusion-v1-5 não foi capaz de gerar imagens de raios-x saudáveis realistas.

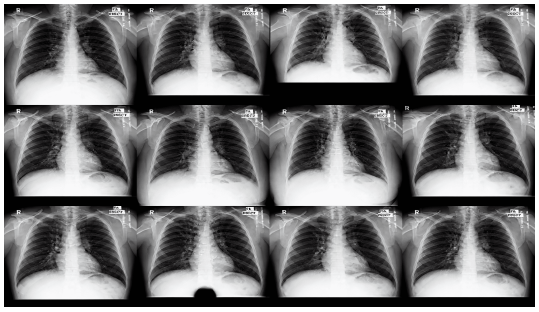
O **Conjunto 2**, do gráfico apresentado na Figura 4.3 (barra da cor verde), representa a avaliação de desempenho do Modelo 01 na geração de imagens de raios-x



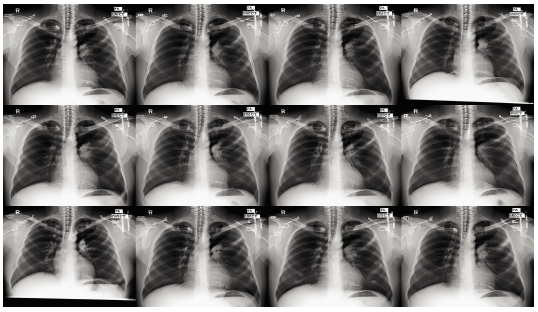
(a) **Conjunto 1**: Conjunto gerado pelo modelo pré-treinado *Foundation Model* stable-diffusion-v1-5 (**F.M**). O modelo foi treinado no conjunto de dados ***laion-aesthetics v2 5+*** [31].



(b) **Conjunto 2**: Conjunto gerado pelo Modelo 01 (**M.1**). O modelo foi treinado utilizando o otimizador **Adam8bit**, utilizando as configurações do Modelo 01 apresentadas na Tabela 3.6.

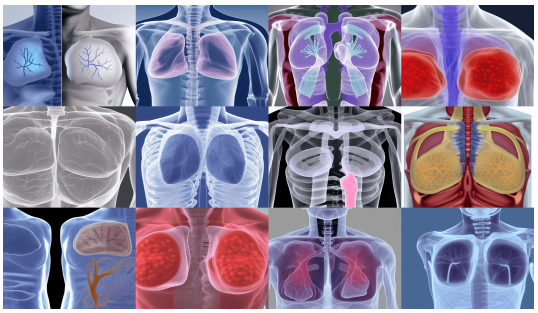


(c) **Conjunto 3**: Conjunto gerado pelo Modelo 02 (**M.2**). O modelo foi treinado utilizando o otimizador **Adam8bit**, utilizando as configurações do Modelo 02 apresentadas na Tabela 3.6.



(d) **Conjunto 4**: Conjunto gerado pelo Modelo 03 (**M.3**). O modelo foi treinado utilizando o otimizador **Adafactor**, utilizando as configurações do Modelo 03 apresentadas na Tabela 3.6.

(e) **Conjunto 5**: Conjunto gerado pelo Modelo 04 (**M.4**). O modelo foi treinado utilizando o otimizador **DAdaptSGD**, utilizando as configurações do Modelo 04 apresentadas na Tabela 3.6.



(f) **Conjunto 6**: Conjunto gerado pelo Modelo 05 (**M.5**). O modelo foi treinado utilizando o otimizador **Prodigy**, utilizando as configurações do Modelo 05 apresentadas na Tabela 3.6.

Figura 4.2: Imagens de raios-x do tórax geradas pelos modelos apresentados na Tabela 3.6 no processo de geração de imagens de raios-x não saudáveis. Todas foram geradas pelos modelos treinados por 20 épocas em um conjunto de dados de 30 imagens de raios-x do tórax. Todas as imagens de raios-x foram geradas utilizando *prompt* "*Human chest x-ray with tuberculosis. Bilateral miliary nodules with Right Middle Lobe infiltrate. Right pleural effusion*".

saudáveis representadas como **Conjunto 2** na Figura 4.1b. Para esse conjunto de imagens, o avaliador avaliou-o como sendo "MUITO Realista". Portanto, podemos afirmar, segundo a avaliação do médico, que o Modelo 01 foi capaz de gerar imagens de raios-x

saudáveis de excelente qualidade. O mesmo processo de apresentação de avaliação de desempenho dos modelos na geração de imagens saudáveis, aplicado aos Conjuntos 1 e 2, também se aplica aos Conjuntos 3, 4, 5 e 6 da Figura 4.3 (barras da cor verde), aos seus correspondentes conjuntos de imagens da Figura 4.1.

O procedimento para a apresentação das avaliações do desempenho dos modelos na geração de imagens de raios-x não saudáveis é idêntico ao procedimento para apresentação das avaliações do desempenho dos modelos na geração de imagens de raios-x saudáveis. Sendo assim, o Conjunto 1, do gráfico da Figura 4.3 (barra da cor marrom), representa a avaliação de desempenho do modelo *Foundation Model stable-diffusion-v1-5* na geração de imagens de raios-x não saudáveis representadas como **Conjunto 1** na Figura 4.2a. Com base na avaliação do conjunto de imagens em questão, o avaliador classificou-o como "Totalmente NÃO Realista". Assim, segundo as avaliações médicas, podemos concluir que o modelo *Foundation Model stable-diffusion-v1-5* não conseguiu gerar imagens realistas de raios-x não saudáveis. É importante salientar que, conforme as avaliações médicas, o modelo *Foundation Model stable-diffusion-v1-5* também não conseguiu gerar imagens realistas de raios-x do tórax saudáveis.

O Conjunto 2, do gráfico da Figura 4.3 (barra da cor marrom), representa a avaliação de desempenho do Modelo 01 na geração de imagens de raios-x não saudáveis representadas como Conjunto 2 na Figura 4.2b. Para esse conjunto de imagens, o avaliador avaliou-o como sendo "RAZOAVELMENTE Realista". Portanto, podemos afirmar, segundo a avaliação do médico, que o Modelo 01 foi capaz de gerar imagens de raios-x não saudáveis de qualidade razoável. O mesmo processo de apresentação de avaliação de desempenho dos modelos na geração de imagens saudáveis, aplicado aos Conjuntos 1 e 2, também se aplica aos Conjuntos 3, 4, 5 e 6 da Figura 4.3 (barra da cor marrom), aos seus correspondentes conjuntos de imagens da Figura 4.2.

Com base no resultado de avaliações realizadas nesta primeira fase, os modelos treinados utilizando o otimizador Adam8bit (Modelo 01 e Modelo 02) apresentaram melhor desempenho na geração de imagens de raios-x realistas. Esses modelos, segundo avaliações realizadas por um médico, foram os que apresentaram melhor desempenho quanto a geração de imagens de raios-x realistas. Esses resultados motivaram a realização de uma série de treinamentos utilizando apenas o otimizador Adam8bit, uma vez que é o que mostrou desempenho satisfatório. Ainda, é importante salientar, que entende-se a limitação de ter-se apenas um avaliador, portanto para a fase 2 foram obtidas duas avaliações de profissionais da saúde.

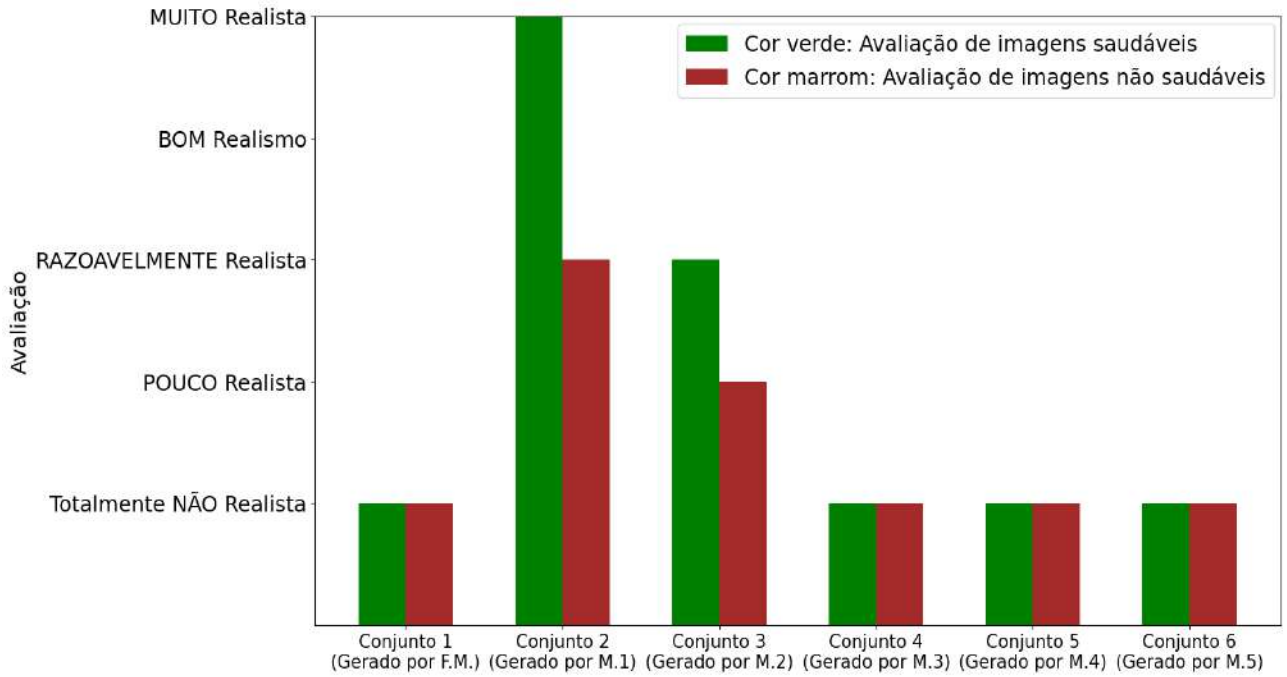


Figura 4.3: Este gráfico ilustra a avaliação de desempenho dos modelos na geração de imagens de raios-x, realizada por um médico. Os conjuntos das imagens são avaliados em uma escala de 1 a 5 no eixo y. Os conjuntos avaliados como “Totalmente NÃO Realistas” são representados na posição 1. Subindo na escala, os conjuntos das imagens considerados “POUCO Realistas” são posicionados em 2. Na posição 3, encontram-se os conjuntos das imagens avaliados como “RAZOAVELMENTE Realistas”. Os conjuntos das imagens que alcançaram um “BOM Realismo” são representados na posição 4. Finalmente, os conjuntos das imagens que foram avaliados como “MUITO Realistas” ocupam a posição 5 do eixo y. As barras na cor verde representam avaliações de imagens saudáveis e as barras na cor marrom representam avaliações de imagens não saudáveis.

4.2 Segunda Fase de Avaliação das Imagens Geradas

As avaliações de desempenho dos modelos na geração de imagens de raios-x, realizadas por um médico na primeira fase, mostrou que modelos treinados utilizando o otimizador Adam8bit apresentaram melhor desempenho. Sendo assim, nesta segunda fase, foram realizados treinamentos utilizando apenas o otimizador Adam8bit (utilizando as configurações do Modelo 1 e Modelo 2 apresentadas na Tabela 3.6). Para isso, preparamos sete conjuntos de dados de 5, 10, 20, 40, 60, 80 e 100 imagens para treinar os modelos. Em cada conjunto, mantivemos uma proporção equilibrada, com 50% de imagens saudáveis e 50% de imagens não saudáveis, todas acompanhadas de seus respectivos

relatórios clínicos. O objetivo é testar o impacto da quantidade de imagens no treino de *Foundation models*. No entanto, devido à falta de especialistas suficientes para avaliar a grande quantidade de imagens geradas nos levou a uma decisão estratégica. Optamos por selecionar apenas as imagens geradas pelos modelos treinados utilizando conjuntos de dados de 20 e 100 imagens. Essa escolha visou tornar o processo de avaliação mais viável.

Para facilitar o processo de avaliação das imagens geradas, conforme mencionado anteriormente, optamos por destacar apenas quatro grupos de imagens. O primeiro grupo é composto por imagens geradas por um modelo treinado por 20 épocas em um conjunto de 20 imagens. O segundo grupo inclui imagens produzidas por um modelo treinado pelo mesmo período, mas em um conjunto maior, de 100 imagens. O terceiro grupo abrange imagens geradas por um modelo treinado por um período mais longo, de 100 épocas, em um conjunto de 20 imagens. Finalmente, o quarto grupo se refere a imagens geradas por um modelo treinado por 100 épocas em um conjunto de 100 imagens. Esse destaque foi aplicado tanto para as imagens saudáveis geradas pelos Modelos 01 e 02, quanto para as imagens não saudáveis, também geradas pelos mesmos modelos.

Para a geração das imagens de raios-x, foram utilizados os seguintes *prompts*: "*Healthy chest x-ray*", para a geração de imagens de raios-x saudáveis; "*Human chest x-ray with large area of cavitation in the right upper lobe*", "*Human chest x-ray with bilateral nodular infiltrate*" e "*Human chest x-ray with right pleural effusion*" (Tabela 4.3) para a geração de imagens de raios-x não saudáveis.

A análise do desempenho dos modelos empregados na geração das imagens radiográficas durante a segunda etapa foi conduzida por duas profissionais médicas. Os critérios de avaliação e os fatores considerados estão detalhados na Tabela 4.4.

Tabela 4.3: *Prompts* utilizados para gerar imagens de raios-x do tórax saudáveis e não saudáveis.

Prompt para geração de imagens saudáveis	*Prompts* para geração de imagens não saudáveis
"*Healthy chest x-ray*"	"*Human chest x-ray with large area of cavitation in the right upper lobe*"
	"*Human chest x-ray with bilateral nodular infiltrate*"
	"*Human chest x-ray with right pleural effusion*"

Tabela 4.4: Escalas e fatores utilizados na avaliação das imagens geradas na segunda fase dos testes.

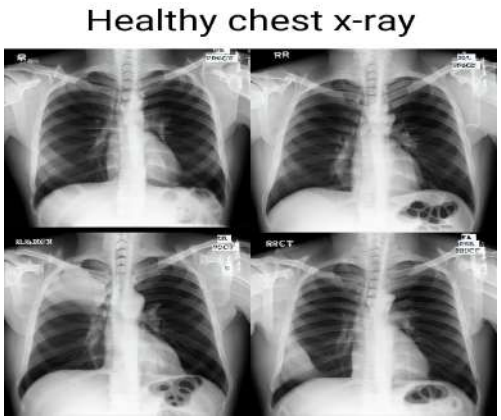
Escala	Fatores	
	De acordo com a Imagem	De acordo com o *PROMPT*
1	NÃO realista	Em desacordo
2	Realista	Concorda em parte
3	MUITO realista	Concorda bastante

4.2.1 Imagens Saudáveis Geradas Pelo Modelo 01 e Avaliações dos Médicos

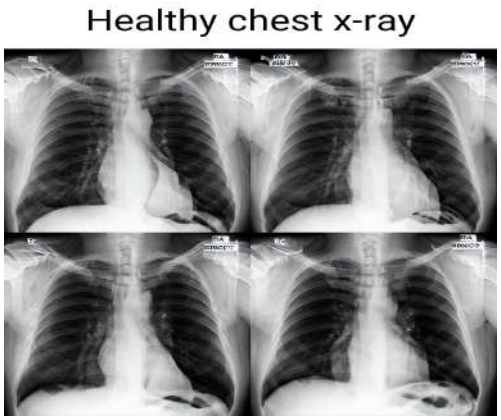
Nesta seção, apresentamos as imagens de raios-x geradas pelo Modelo 1 (Figura 4.4), bem como as avaliações dos profissionais médicos sobre o desempenho desse modelo na geração das imagens de raios-x. Cada figura contém quatro grupos de imagens geradas por um modelo, juntamente com dois gráficos de avaliação. Esses gráficos consideram o realismo das imagens e a conformidade das imagens de raios-x geradas com os *prompts* utilizados para gerá-las. Por exemplo, o Gráfico 4.4e, mostra as avaliações do desempenho do Modelo 01 na geração de imagens de raios-x saudáveis em termos de realismo. Já o Gráfico 4.4f, representa as avaliações de desempenho do Modelo 01 na geração de imagens de raios-x saudáveis de acordo com o *prompt*.

4.2.2 Imagens Saudáveis Geradas Pelo Modelo 02 e Avaliações dos Médicos

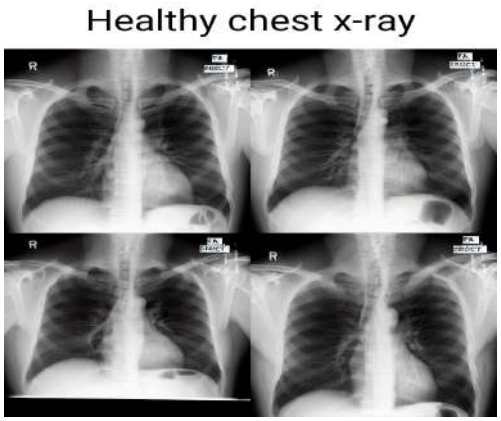
Nesta seção, novamente, apresentamos as imagens de raios-x geradas pelo Modelo 2 (Figura 4.5), bem como as avaliações dos profissionais médicos sobre o desempenho desse modelo na geração das imagens de raios-x. Cada figura contém quatro grupos de imagens geradas por um modelo, juntamente com dois gráficos de avaliação. Esses gráficos consideram o realismo das imagens e a conformidade das imagens de raios-x geradas com os *prompts* utilizados para gerá-las. Por exemplo, o Gráfico 4.5e, mostra as avaliações do desempenho do Modelo 02 na geração de imagens de raios-x saudáveis em



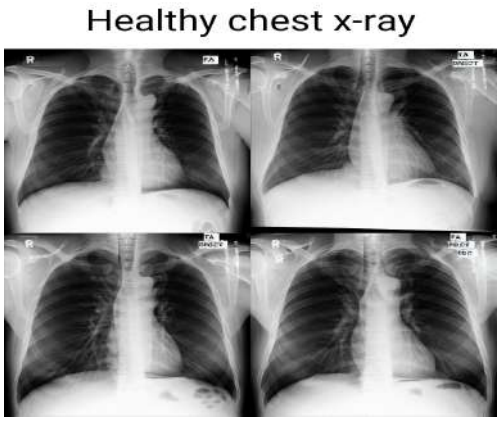
(a) **Conjunto 1**: Conjunto gerado pelo Modelo 01 (**M.1**). O modelo foi treinado por 20 épocas, num conjunto de dados de 20 imagens.



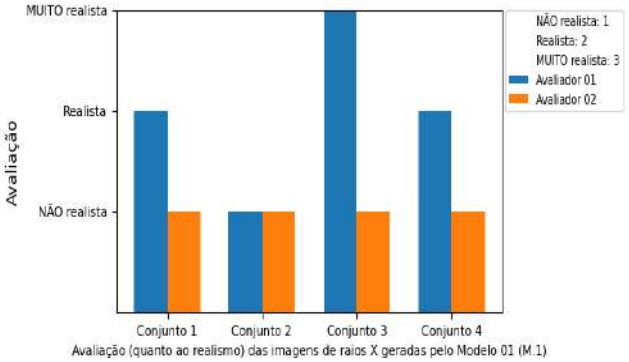
(b) **Conjunto 2**: Conjunto gerado pelo Modelo 01 (**M.1**). O modelo foi treinado por 100 épocas, num conjunto de dados de 20 imagens.



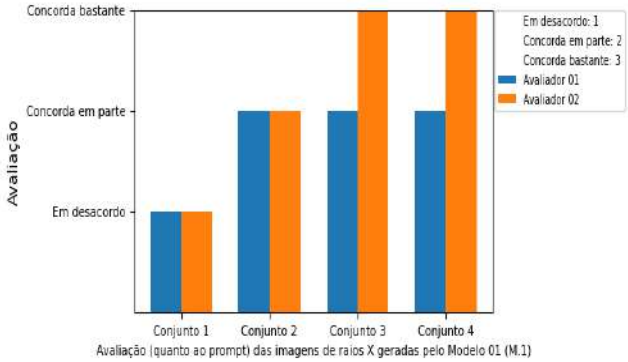
(c) **Conjunto 3**: Conjunto gerado pelo Modelo 01 (**M.1**). O modelo foi treinado por 20 épocas, num conjunto de dados de 100 imagens.



(d) **Conjunto 4**: Conjunto gerado pelo Modelo 01 (**M.1**). O modelo foi treinado por 100 épocas, num conjunto de dados de 100 imagens.



(e) Avaliação feita por DUAS especialistas quanto ao desempenho do Modelo 01 na geração de imagens de raios-x saudáveis, quanto ao realismo.



(f) Avaliação feita por DUAS especialistas quanto ao desempenho do Modelo 01 na geração de imagens de raios-x saudáveis, de acordo com o *prompt*.

Figura 4.4: Imagens de raios-x do tórax geradas pelo Modelo 01 (M.1) e gráficos de avaliação de desempenho dos modelos. Os gráficos (e) e (f) representam avaliações de desempenho do Modelo 01 na geração de imagens de raios-x realistas e saudáveis. A Avaliação foi realizada considerando o realismo das imagens de raios-x (gráfico (e)) e de acordo com *prompt* utilizando (gráfico (f)). No topo de cada conjunto de imagens (**a**, **b**, **c** e **d**) constam os *prompts* utilizados. Todos os conjuntos foram gerados utilizando as configurações do Modelo 01 apresentadas na Tabela 3.6.

termos de realismo. Já o Gráfico 4.5f, representa as avaliações de desempenho do Modelo 02 na geração de imagens de raios-x saudáveis de acordo com o *prompt*.

4.2.3 Imagens Não Saudáveis Geradas Pelo Modelo 01 e Avaliações dos Médicos

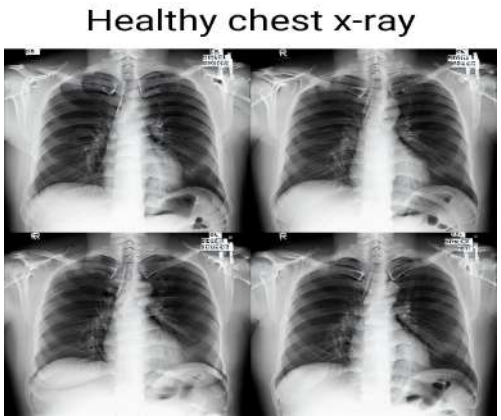
Nesta seção, apresentamos uma série de imagens de raios-x geradas pelos submodelos do Modelo 1 (Figuras 4.6, 4.7 e 4.8), bem como as avaliações dos profissionais médicos sobre o desempenho desses submodelos na geração das imagens de raios-x. Cada figura contém quatro conjuntos de imagens geradas pelos submodelos, juntamente com dois gráficos de avaliação. Esses gráficos consideram o realismo das imagens e a conformidade das imagens de raios-x geradas com os *prompts* utilizados para gerá-las. Por exemplo, o Gráfico 4.6e, mostra as avaliações do desempenho dos submodelos do Modelo 01 na geração de imagens de raios-x não saudáveis em termos de realismo. Já o Gráfico 4.6f, representa as avaliações de desempenho dos submodelos do Modelo 01 na geração de imagens de raios-x não saudáveis, de acordo com o *prompt*. Nesses gráficos, é possível visualizar e comparar a visão das avaliadoras quanto ao realismo das imagens geradas e se as imagens geradas correspondem aos *prompts* utilizados para gerá-las.

No topo de cada conjunto de imagens apresentadas nas Figuras 4.6, 4.7 e 4.8, constam os *prompts* utilizados para gerar o conjunto de imagens em causa.

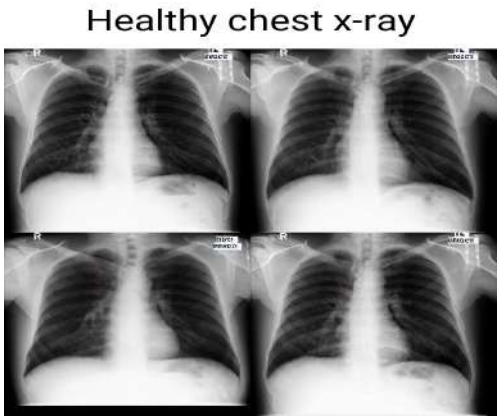
4.2.4 Imagens Não Saudáveis Geradas Pelo Modelo 02 e Avaliações dos Médicos

Nesta seção, apresentamos uma série de imagens de raios-x não saudáveis geradas pelos submodelos do Modelo 2 (Figuras 4.9, 4.10 e 4.11), bem como as avaliações dos profissionais médicos sobre o desempenho desses submodelos na geração das imagens de raios-x. Cada figura contém quatro conjuntos de imagens geradas pelos submodelos, juntamente com dois gráficos de avaliação. Esses gráficos consideram o realismo das imagens e a conformidade das imagens de raios-x geradas com os *prompts* utilizados para gerá-las. Por exemplo, o Gráfico 4.9e, mostra as avaliações do desempenho dos submodelos do Modelo 02 na geração de imagens de raios-x não saudáveis em termos de realismo. Já o Gráfico 4.9f, representa as avaliações de desempenho dos submodelos do Modelo 02 na geração de imagens de raios-x não saudáveis, de acordo com o *prompt*. Nesses gráficos, novamente, é possível visualizar e comparar a visão das avaliadoras quanto ao realismo das imagens geradas e se as imagens geradas correspondem aos *prompts* utilizados para gerá-las.

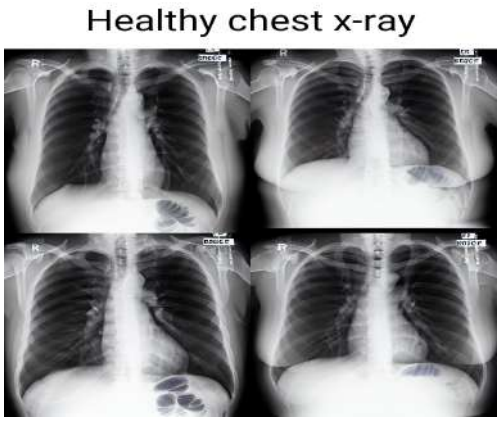
No topo de cada conjunto de imagens apresentadas nas Figuras 4.9, 4.10 e 4.11, constam os *prompts* utilizados para gerar os conjuntos de imagens em causa.



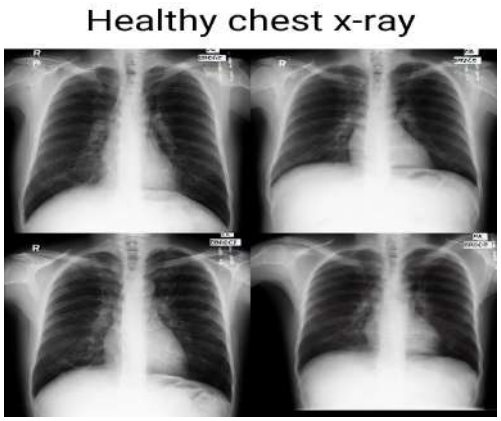
(a) **Conjunto 1**: Conjunto gerado pelo Modelo 02 (**M.2**). O modelo foi treinado por 20 épocas, num conjunto de dados de 20 imagens.



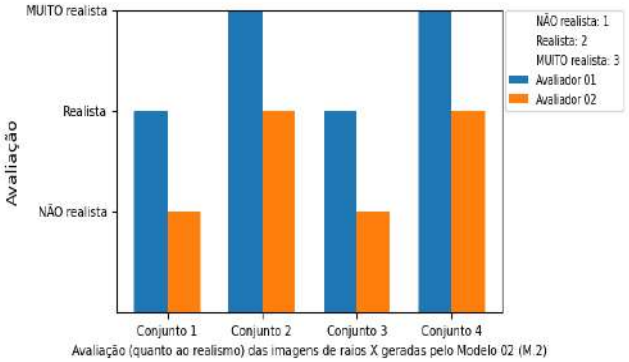
(b) **Conjunto 2**: Conjunto gerado pelo Modelo 02 (**M.2**). O modelo foi treinado por 100 épocas, num conjunto de dados de 20 imagens.



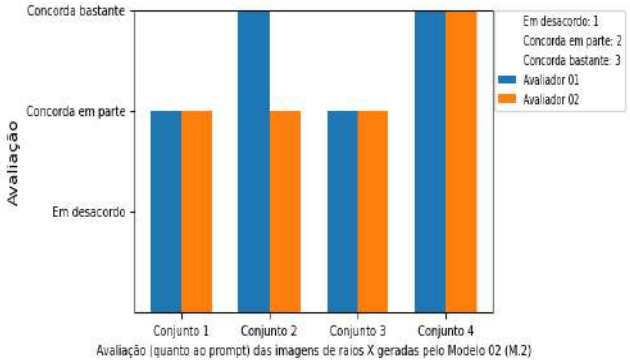
(c) **Conjunto 3**: Conjunto gerado pelo Modelo 02 (**M.2**). O modelo foi treinado por 20 épocas, num conjunto de dados de 100 imagens.



(d) **Conjunto 4**: Conjunto gerado pelo Modelo 02 (**M.2**). O modelo foi treinado por 100 épocas, num conjunto de dados de 100 imagens.

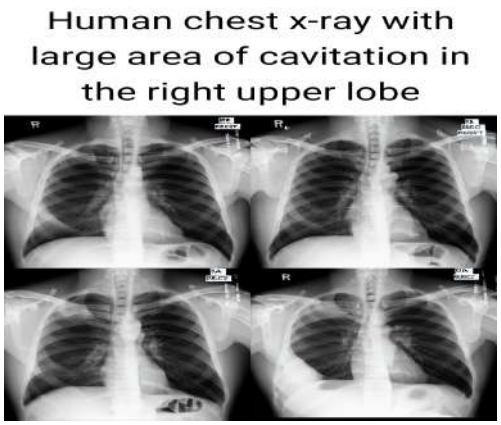


(e) Avaliação feita por DUAS especialistas quanto ao desempenho do Modelo 02 na geração de imagens de raios-x saudáveis, quanto ao realismo.

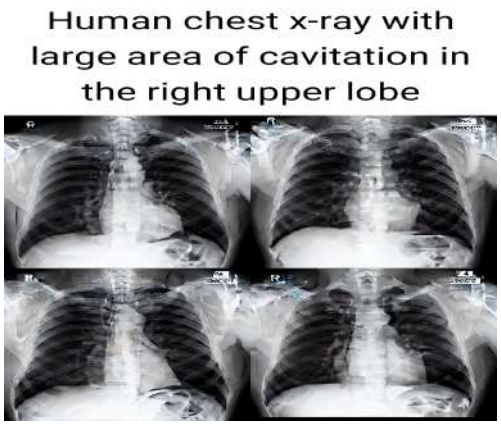


(f) Avaliação feita por DUAS especialistas quanto ao desempenho do Modelo 02 na geração de imagens de raios-x saudáveis, de acordo com o *prompt*.

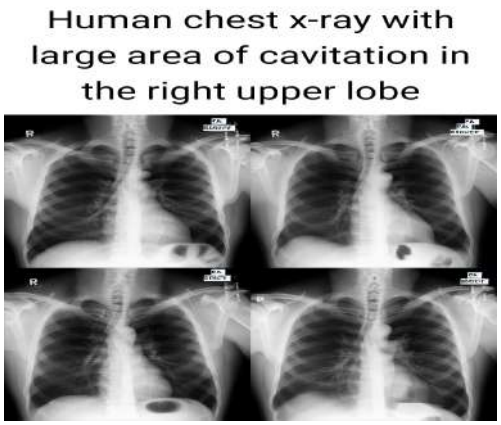
Figura 4.5: Imagens de raios-x do tórax geradas pelo Modelo 02 (**M.2**) e gráficos de avaliação de desempenho dos modelos. Os gráficos (e) e (f) representam avaliações de desempenho do Modelo 02 na geração de imagens de raios-x realistas e saudáveis. A Avaliação foi realizada considerando o realismo das imagens de raios-x (gráfico (e)) e de acordo com *prompt* utilizando (gráfico (f)). No topo de cada conjunto de imagens (**a**, **b**, **c** e **d**) constam os *prompts* utilizados. Todos os conjuntos foram gerados utilizando as configurações do Modelo 02 apresentadas na Tabela 3.6.



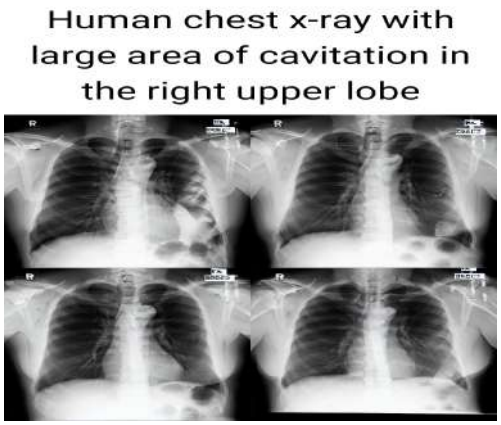
(a) **Conjunto 1**: Conjunto gerado pelo Modelo 01 (**M.1**). O modelo foi treinado por 20 épocas, num conjunto de dados de 20 imagens.



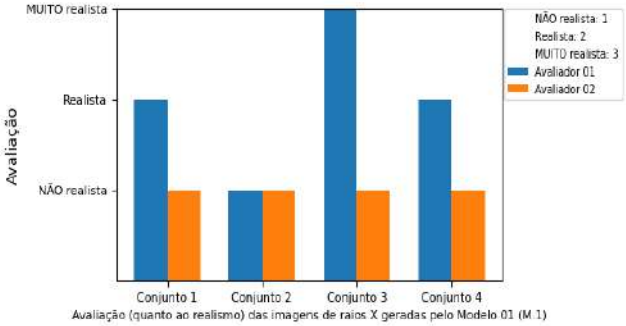
(b) **Conjunto 2**: Conjunto gerado pelo Modelo 01 (**M.1**). O modelo foi treinado por 100 épocas, num conjunto de dados de 20 imagens.



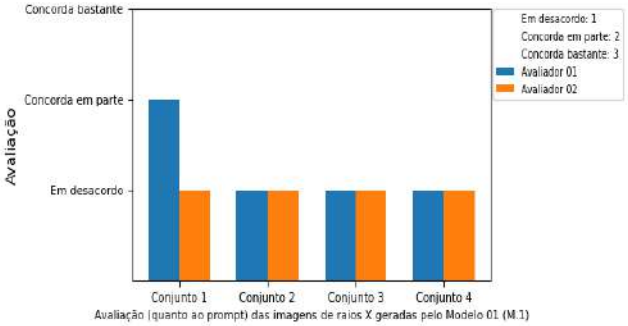
(c) **Conjunto 3**: Conjunto gerado pelo Modelo 01 (**M.1**). O modelo foi treinado por 20 épocas, num conjunto de dados de 100 imagens.



(d) **Conjunto 4**: Conjunto gerado pelo Modelo 01 (**M.1**). O modelo foi treinado por 100 épocas, num conjunto de dados de 100 imagens.

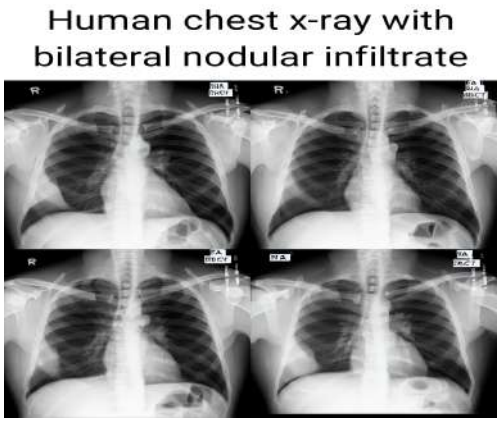


(e) Avaliação feita por DUAS especialistas quanto ao desempenho do Modelo 01 na geração de imagens de raios-x não saudáveis, quanto ao realismo.

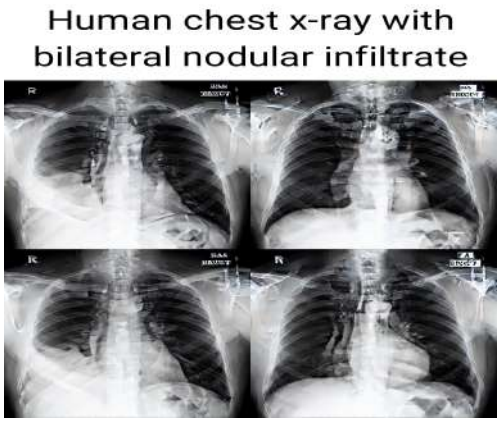


(f) Avaliação feita por DUAS especialistas quanto ao desempenho do Modelo 01 na geração de imagens de raios-x não saudáveis, de acordo com o *prompt*.

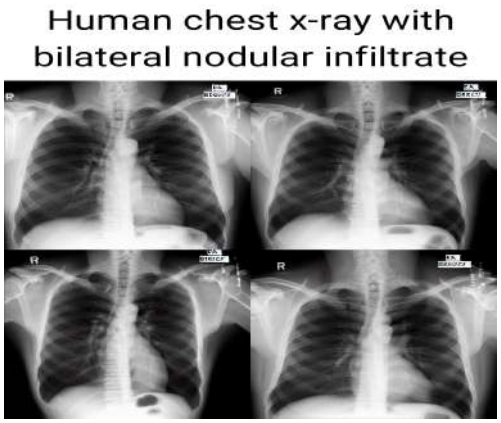
Figura 4.6: Imagens de raios-x do tórax geradas pelo Modelo 01 (M.1) e os gráficos de avaliação de desempenho dos modelos. Os gráficos (e) e (f) representam avaliações de desempenho do Modelo 01 na geração de imagens de raios-x realistas e não saudáveis. A Avaliação foi realizada considerando o realismo das imagens de raios-x (gráfico (e)) e de acordo com *prompt* utilizando (gráfico (f)). No topo de cada conjunto de imagens (**a**, **b**, **c** e **d**) constam os *prompts* utilizados. Todos os conjuntos foram gerados utilizando as configurações do Modelo 01 apresentadas na Tabela 3.6.



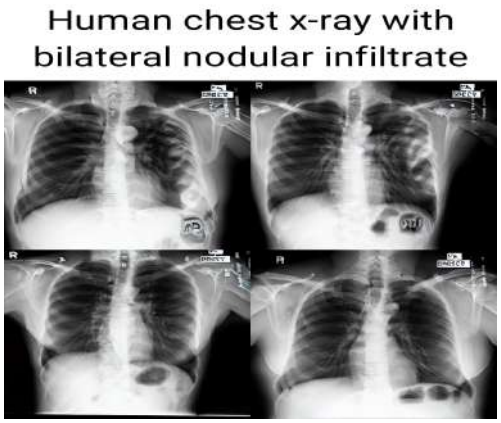
(a) **Conjunto 1**: Conjunto gerado pelo Modelo 01 (**M.1**). O modelo foi treinado por 20 épocas, num conjunto de dados de 20 imagens.



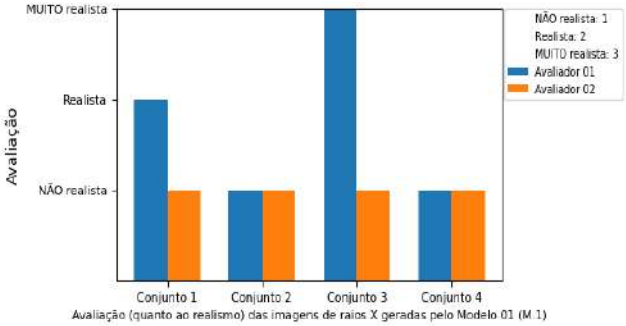
(b) **Conjunto 2**: Conjunto gerado pelo Modelo 01 (**M.1**). O modelo foi treinado por 100 épocas, num conjunto de dados de 20 imagens.



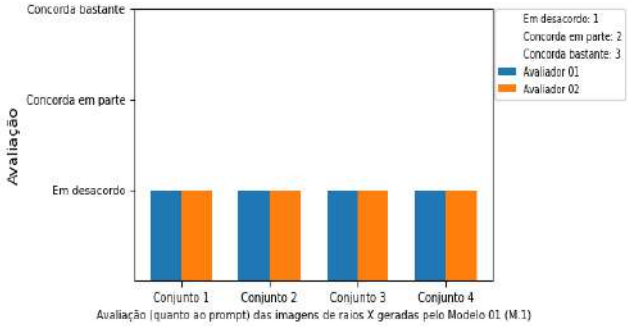
(c) **Conjunto 3**: Conjunto gerado pelo Modelo 01 (**M.1**). O modelo foi treinado por 20 épocas, num conjunto de dados de 100 imagens.



(d) **Conjunto 4**: Conjunto gerado pelo Modelo 01 (**M.1**). O modelo foi treinado por 100 épocas, num conjunto de dados de 100 imagens.

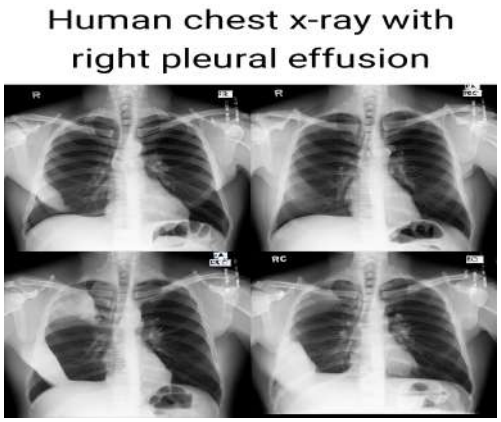


(e) Avaliação feita por DUAS especialistas quanto ao desempenho do Modelo 01 na geração de imagens de raios-x não saudáveis, quanto ao realismo.

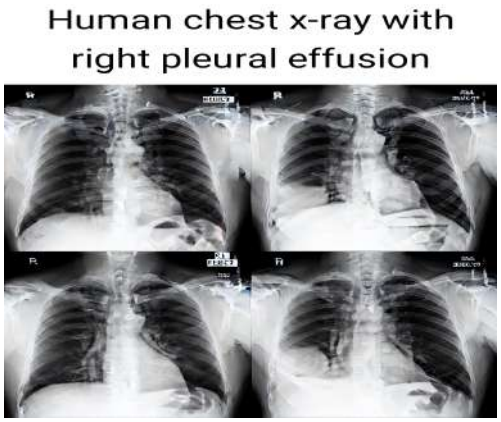


(f) Avaliação feita por DUAS especialistas quanto ao desempenho do Modelo 01 na geração de imagens de raios-x não saudáveis, de acordo com o *prompt*.

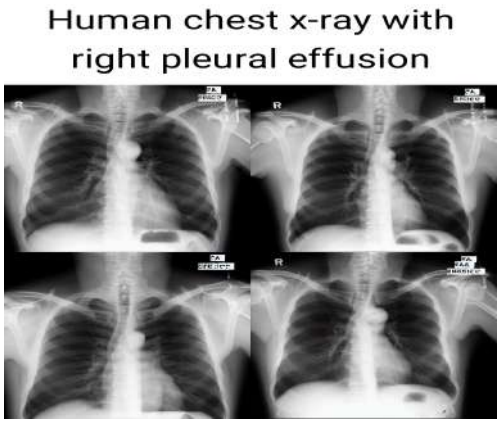
Figura 4.7: Imagens de raios-x do tórax geradas pelo Modelo 01 (**M.1**) e gráficos de avaliação de desempenho dos modelos. Os gráficos (e) e (f) representam avaliações de desempenho do Modelo 01 na geração de imagens de raios-x realistas e não saudáveis. A Avaliação foi realizada considerando o realismo das imagens de raios-x, gráfico (e) e de acordo com *prompt* utilizando, gráfico (f). No topo de cada conjunto de imagens (**a**, **b**, **c** e **d**) constam os *prompts* utilizados. Todos os conjuntos foram gerados utilizando as configurações do Modelo 01 apresentadas na Tabela 3.6.



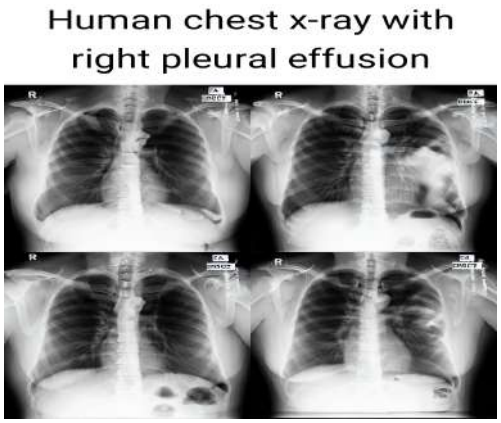
(a) **Conjunto 1**: Conjunto gerado pelo Modelo 01 (**M.1**). O modelo foi treinado por 20 épocas, num conjunto de dados de 20 imagens.



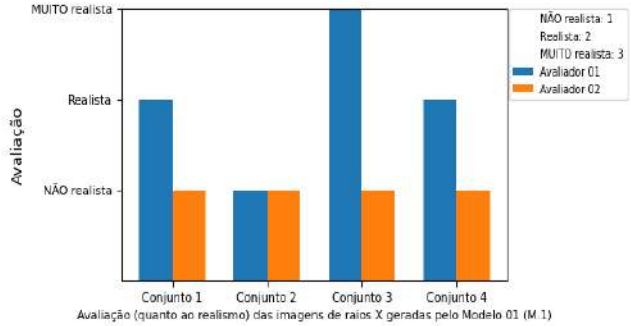
(b) **Conjunto 2**: Conjunto gerado pelo Modelo 01 (**M.1**). O modelo foi treinado por 100 épocas, num conjunto de dados de 20 imagens.



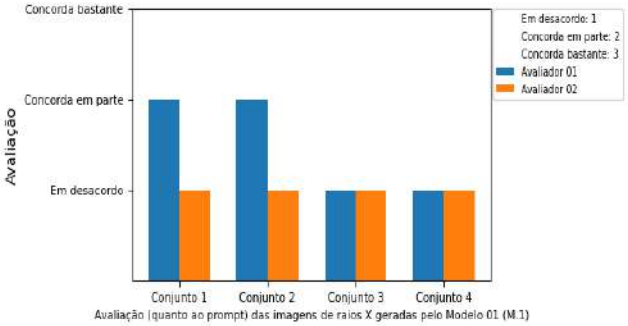
(c) **Conjunto 3**: Conjunto gerado pelo Modelo 01 (**M.1**). O modelo foi treinado por 20 épocas, num conjunto de dados de 100 imagens



(d) **Conjunto 4**: Conjunto gerado pelo Modelo 01 (**M.1**). O modelo foi treinado por 100 épocas, num conjunto de dados de 100 imagens.



(e) Avaliação feita por DUAS especialistas quanto ao desempenho do Modelo 01 na geração de imagens de raios-x não saudáveis, quanto ao realismo.

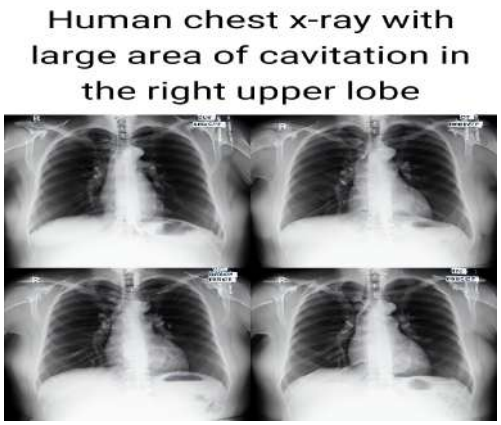


(f) Avaliação feita por DUAS especialistas quanto ao desempenho do Modelo 01 na geração de imagens de raios-x não saudáveis, de acordo com o *prompt*.

Figura 4.8: Imagens de raios-x do tórax geradas pelo Modelo 01 (M.1) e gráficos de avaliação de desempenho dos modelos. Os gráficos (e) e (f) representam avaliações de desempenho do Modelo 01 na geração de imagens de raios-x realistas e não saudáveis. A Avaliação foi realizada considerando o realismo das imagens de raios-x (gráfico (e)) e de acordo com *prompt* utilizando (gráfico (f)). No topo de cada conjunto de imagens (**a**, **b**, **c** e **d**) constam os *prompts* utilizados. Todos os conjuntos foram gerados utilizando as configurações do Modelo 01 apresentadas na Tabela 3.6.

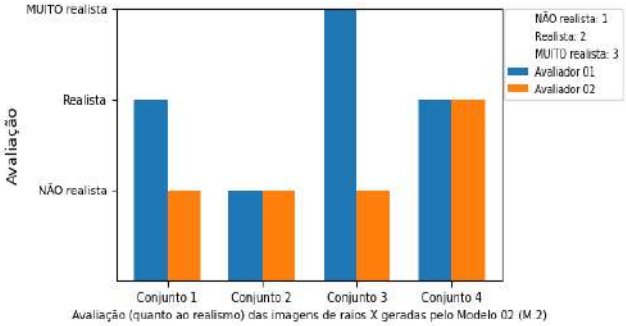
(a) **Conjunto 1**: Conjunto gerado pelo Modelo 02 (**M.2**). O modelo foi treinado por 20 épocas, num conjunto de dados de 20 imagens.

(b) **Conjunto 2**: Conjunto gerado pelo Modelo 02 (**M.2**). O modelo foi treinado por 100 épocas, num conjunto de dados de 20 imagens.

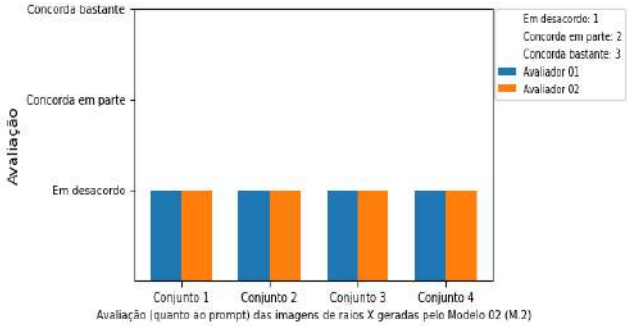


(c) **Conjunto 3**: Conjunto gerado pelo Modelo 02 (**M.2**). O modelo foi treinado por 20 épocas, num conjunto de dados de 100 imagens.

(d) **Conjunto 4**: Conjunto gerado pelo Modelo 02 (**M.2**). O modelo foi treinado por 100 épocas, num conjunto de dados de 100 imagens.

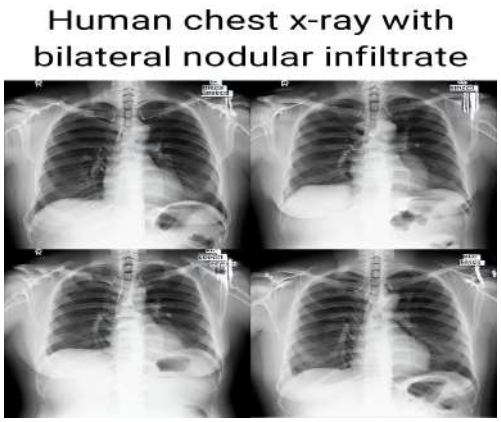


(e) Avaliação feita por DUAS especialistas quanto ao desempenho do Modelo 02 na geração de imagens de raios-x não saudáveis, quanto ao realismo.

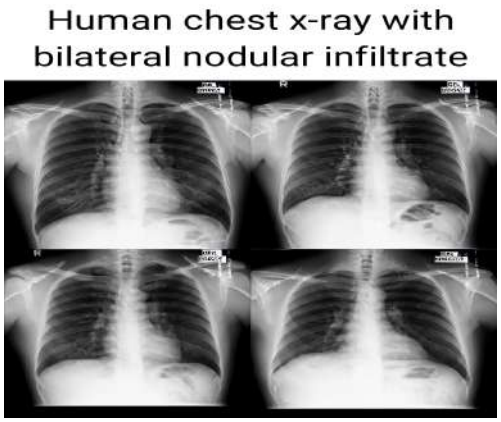


(f) Avaliação feita por DUAS especialistas quanto ao desempenho do Modelo 02 na geração de imagens de raios-x não saudáveis, de acordo com o *prompt*.

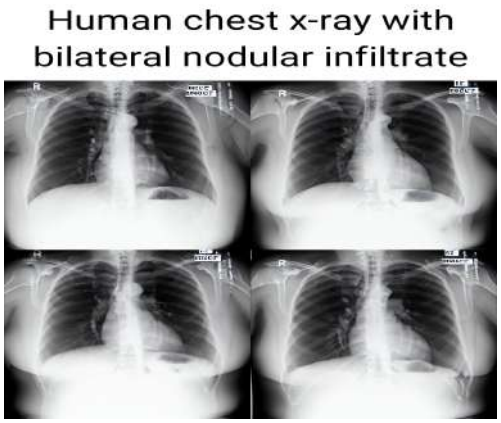
Figura 4.9: As quatro imagens de raios-x do tórax foram gerados pelo Modelo 02. Os gráficos (e) e (f) representam avaliações de desempenho do Modelo 02 na geração de imagens de raios-x realistas e não saudáveis. A Avaliação foi realizada considerando o realismo das imagens de raios-x, gráfico (e), e de acordo com *prompt* utilizando, gráfico (f). No topo de cada conjunto de imagens (**a**, **b**, **c** e **d**) constam os *prompts* utilizados. Todos os conjuntos foram gerados utilizando as configurações do Modelo 02 apresentadas na Tabela 3.6.



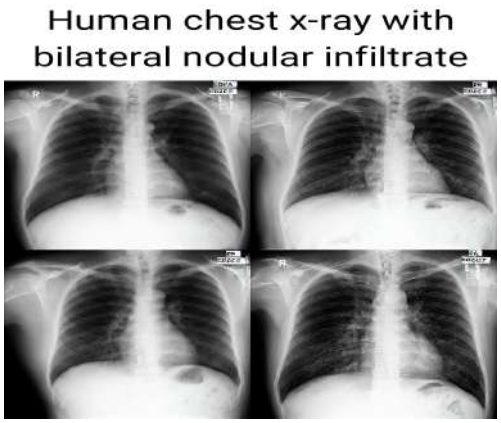
(a) **Conjunto 1**: Conjunto gerado pelo Modelo 02 (**M.2**). O modelo foi treinado por 20 épocas, num conjunto de dados de 20 imagens.



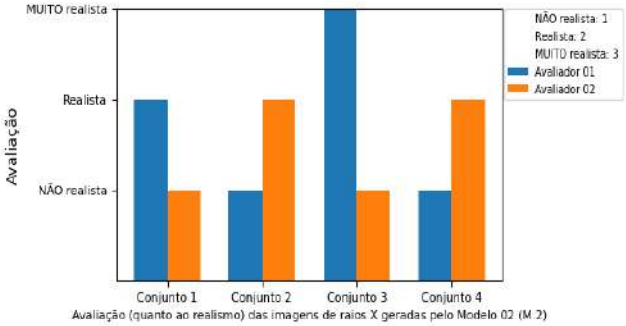
(b) **Conjunto 2**: Conjunto gerado pelo Modelo 02 (**M.2**). O modelo foi treinado por 100 épocas, num conjunto de dados de 20 imagens.



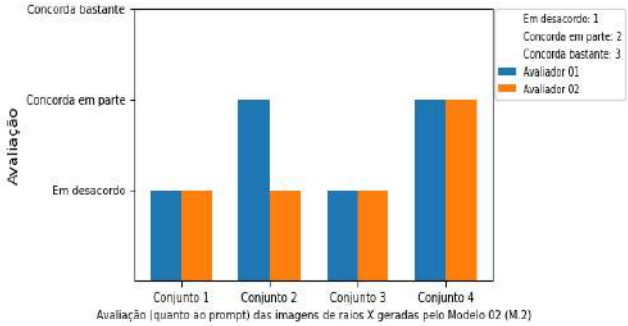
(c) **Conjunto 3**: Conjunto gerado pelo Modelo 02 **(M.2)**. O modelo foi treinado por 20 épocas, num conjunto de dados de 100 imagens.



(d) **Conjunto 4**: Conjunto gerado pelo Modelo 02 (**M.2**). O modelo foi treinado por 100 épocas, num conjunto de dados de 100 imagens.

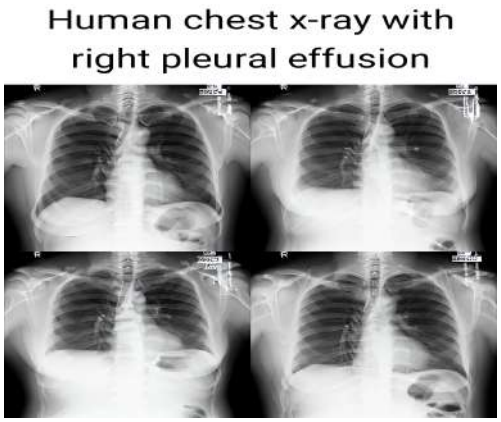


(e) Avaliação feita por DUAS especialistas quanto ao desempenho do Modelo 02 na geração de imagens de raios-x não saudáveis, quanto ao realismo.

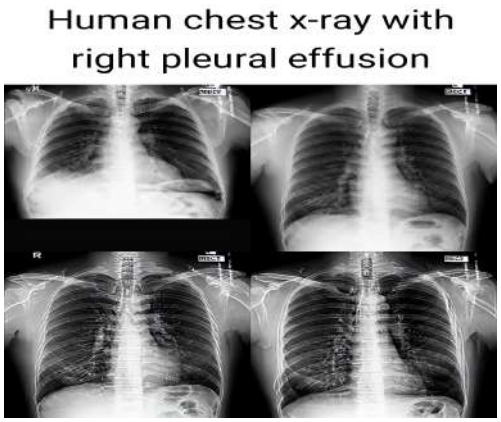


(f) Avaliação feita por DUAS especialistas quanto ao desempenho do Modelo 02 na geração de imagens de raios-x não saudáveis, de acordo com o *prompt*.

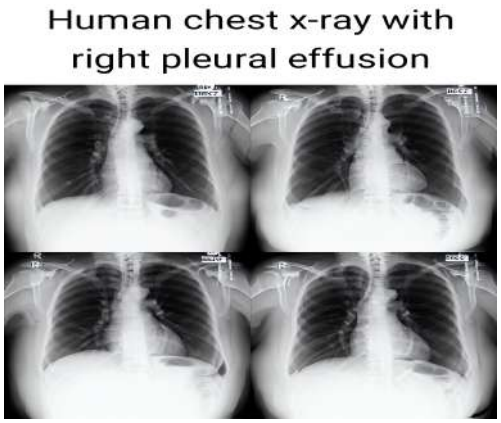
Figura 4.10: Imagens de raios-x do tórax geradas pelo Modelo 02 (M.2) e gráficos de avaliação de desempenho dos modelos. Os gráficos (e) e (f) representam avaliações de desempenho do Modelo 02 na geração de imagens de raios-x realistas e não saudáveis. A Avaliação foi realizada considerando o realismo das imagens de raios-x (gráfico (e)) e de acordo com *prompt* utilizando (gráfico (f)). No topo de cada conjunto de imagens (**a**, **b**, **c** e **d**) constam os *prompts* utilizados. Todos os conjuntos foram gerados utilizando as configurações do Modelo 02 apresentadas na Tabela 3.6.



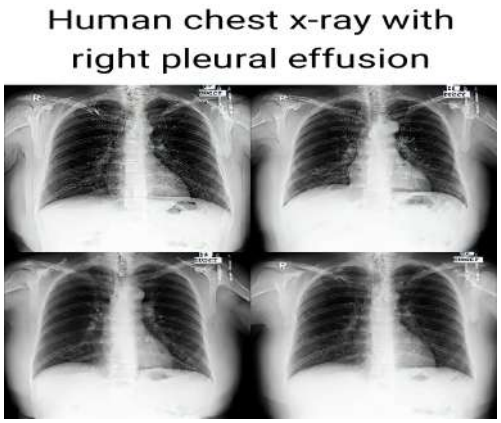
(a) **Conjunto 1**: Conjunto gerado pelo Modelo 02 (**M.2**). O modelo foi treinado por 20 épocas, num conjunto de dados de 20 imagens.



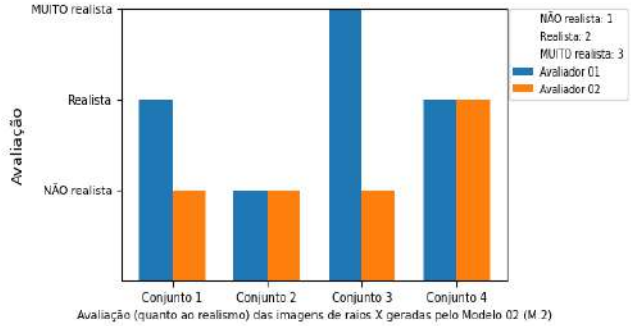
(b) **Conjunto 2**: Conjunto gerado pelo Modelo 02 (**M.2**). O modelo foi treinado por 100 épocas, num conjunto de dados de 20 imagens.



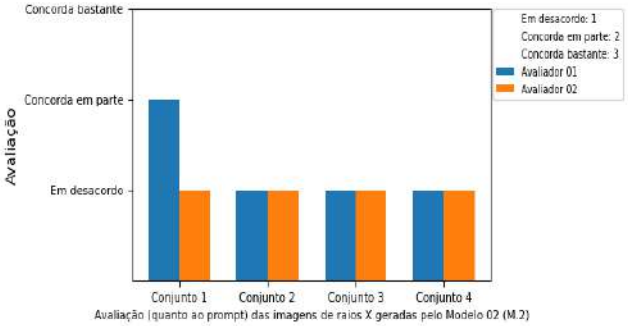
(c) **Conjunto 3**: Conjunto gerado pelo Modelo 02 (**M.2**). O modelo foi treinado por 20 épocas, num conjunto de dados de 100 imagens.



(d) **Conjunto 4**: Conjunto gerado pelo Modelo 02 (**M.2**). O modelo foi treinado por 100 épocas, num conjunto de dados de 100 imagens.



(e) Avaliação feita por DUAS especialistas quanto ao desempenho do Modelo 02 na geração de imagens de raios-x não saudáveis, quanto ao realismo.



(f) Avaliação feita por DUAS especialistas quanto ao desempenho do Modelo 02 na geração de imagens de raios-x não saudáveis, de acordo com o *prompt*.

Figura 4.11: Imagens de raios-x do tórax geradas pelo Modelo 02 e gráficos de avaliação de desempenho dos modelos. Os gráficos (e) e (f) representam avaliações de desempenho do Modelo 02 na geração de imagens de raios-x realistas e não saudáveis. A Avaliação foi realizada considerando o realismo das imagens de raios-x (gráfico (e)) e de acordo com *prompt* utilizando (gráfico (f)). No topo de cada conjunto de imagens (**a**, **b**, **c** e **d**) constam os *prompts* utilizados. Todos os conjuntos foram gerados utilizando as configurações do Modelo 02 apresentadas na Tabela 3.6.

4.3 Discussão Geral Sobre os Resultados

Nesta seção, abordamos os resultados das avaliações realizadas pelos médicos em duas fases distintas, primeira e na segunda fase. A seção é estruturada em três partes para facilitar a compreensão. Inicialmente, focamos na primeira fase do estudo, onde discutimos a percepção do avaliador sobre o desempenho dos modelos na geração de imagens de raios-x do tórax. Em seguida, na segunda parte, mudamos nosso foco para a segunda fase da avaliação. Aqui, apresentamos as opiniões de duas avaliadoras sobre a desempenho dos modelos na geração de imagens de raios-x do tórax. Finalmente, na terceira e última parte, fizemos uma comparação entre a qualidade das imagens saudáveis e não saudáveis geradas pelos modelos.

4.3.1 Visão do avaliador sobre o desempenho dos modelos na geração de imagens de raios-x do tórax na primeira fase

As avaliações do desempenho dos modelos na geração de imagens de raios-x do tórax realizadas na primeira fase foram conduzidas por um único médico. Nessa etapa, dos seis modelos usados para gerar as imagens de raios-x, incluindo o *Foundation Model stable-diffusion-v1-5*, apenas o Modelo 01 e o Modelo 02 tiveram um desempenho considerado plausível (Figura 4.3). Estes modelos, de acordo com avaliações médicas, foram capazes de gerar imagens que variam de "Realista" a "MUITO realista".

4.3.2 Visão das avaliadoras sobre o desempenho dos modelos na geração de imagens de raios-x do tórax na segunda fase

As avaliações do desempenho dos modelos na geração de imagens de raios-x do tórax realizadas na segunda fase foram conduzidas por duas médicas. Nessa fase, os modelos 01 e 02, que haviam apresentado melhor desempenho na primeira fase, foram os únicos treinados e utilizados para gerar as imagens de raios-x. Para apresentar uma discussão sobre a visão das avaliadoras, dividimos esta subseção em duas partes: na primeira parte, apresentamos de forma resumida a visão das avaliadoras sobre o desempenho dos modelos na geração de imagens de raios-x saudáveis e na segunda parte, apresentamos também de forma resumida a visão das avaliadoras sobre o desempenho dos modelos na geração de imagens de raios-x não saudáveis. É importante frisar que para a segunda parte, apresentamos também a visão das avaliadoras sobre o desempe-

nho dos modelos na geração de imagens que estejam de acordo com os *prompts* utilizados para gerá-las.

Primeira parte: desempenho dos modelos na geração de imagens de raios-x saudáveis

Nas avaliações dos conjuntos de imagens gerados pelo Modelo 01, as avaliadoras classificaram a maioria dos conjuntos como sendo "Não Realista"ou "Realista"em relação ao seu nível de realismo. Nessa avaliação, houve uma única avaliação (discrepância significativa) onde um avaliador atribui a classificação de "MUITO realista"enquanto o outro avaliador atribui a classificação de "NÃO realista"ao mesmo conjunto de imagens (Conjunto 3, Figura 4.4c). Além disso, quanto à concordância entre as imagens geradas e os *prompts* utilizados para gerá-las, as avaliadoras também foram quase unânimes em "Concordar em parte"ou em "Concordar bastante"que as imagens geradas estão de acordo com os *prompts* utilizados (Conjuntos 2, 3 e 4, Figura 4.4f).

No caso das avaliações sobre o desempenho do Modelo 02, as avaliações realizadas sobre os conjuntos de imagens de raios-x gerados por esse modelo mostram que a maioria das classificações atribuídas pelas avaliadoras varia de "Realista"a "MUITO realista". Notavelmente, dois conjuntos (Conjunto 2 e Conjunto 4, Figura 4.5e) receberam a classificação de "MUITO realista", conforme as avaliações realizadas pelo Avaliador 1, superando a única classificação de "MUITO realista"atribuída pelo mesmo avaliador ao Conjunto 3, gerado pelo Modelo 01. Quanto à concordância entre as imagens geradas e os *prompts* utilizados para gerá-las, as avaliadoras foram quase unânimes em "Concordar em parte"ou em "Concordar bastante"que as imagens geradas pelo Modelo 02 estão de acordo com os *prompts* utilizados (conjuntos 1 a 4, Figura 4.5e). A diferença aqui é que não houve uma única classificação "Em desacordo"como ocorreu na avaliação realizada sobre os conjuntos de imagens gerados pelo Modelo 01.

Segunda parte: desempenho dos modelos na geração de imagens de raios-x não saudáveis

Os conjuntos de imagens gerados pelo Modelo 01 foram submetidos a uma avaliação criteriosa quanto ao realismo. Em três ocasiões distintas, esses conjuntos foram categorizados como "MUITO realistas". Tal classificação foi conferida ao Conjunto 3, que se originou de cada um dos submodelos utilizados na geração das imagens não saudáveis. É importante enfatizar que todas as avaliações que resultaram na classificação de "MUITO realista"foram conduzidas pelo Avaliador 1. Adicionalmente, uma outra parcela dos conjuntos de imagens gerados pelo Modelo 01 foram avaliados quanto ao seu realismo e, em cinco ocasiões distintas, receberam a classificação de "Realista"(conjuntos 1 e 4, Figura 4.6e; Conjunto 1, Figura 4.7f e conjuntos 1 e 4 da Figura 4.8e). É importante destacar que todas as avaliações que resultaram nas classificações de "Realista"e "MUITO realista"para

os conjuntos de imagens gerados pelo Modelo 01 foram realizadas pelo Avaliador 1. Por outro lado, o Avaliador 2 classificou todos os conjuntos de imagens gerados pelo Modelo 01 como sendo "NÃO realistas".

Os conjuntos de imagens gerados pelo Modelo 02 também foram submetidos a uma avaliação criteriosa quanto ao realismo e, em três ocasiões distintas, esses conjuntos foram categorizados como "MUITO realistas". Tal classificação foi conferida ao Conjunto 3, que se originou de cada um dos submodelos utilizados na geração das imagens não saudáveis. É importante enfatizar que todas as avaliações que resultaram na classificação de "MUITO realista" foram conduzidas pelo Avaliador 1. Adicionalmente, uma outra parcela dos conjuntos de imagens gerados pelo Modelo 02 foram avaliados quanto ao realismo e, em nove ocasiões distintas, receberam a classificação de "Realista" (Conjunto 1 e Conjunto 4 da Figura 4.9e; Conjunto 1, 2 e 4 da Figura 4.10e; Conjunto 1 e Conjunto 4 da Figura 4.11f).

É importante destacar que houve consenso entre as avaliadoras ao classificar os conjuntos 4 das figuras 4.9 e 4.11 como sendo "Realista".

Em relação à conformidade dos conjuntos das imagens de raios-x gerados com os *prompts* utilizados para gerá-los, a perspectiva das avaliadoras pode ser expressa da seguinte maneira:

- Modelo 01: As avaliadoras "Concordaram em parte"em três ocasiões que os conjuntos de imagens não saudáveis gerados pelo Modelo 01 estão alinhados com os *prompts* utilizados para gerá-los (Conjunto 1 da Figura 4.6f, conjuntos 1 e 2 da Figura 4.8f). Todas essas classificações foram atribuídas pelo Avaliador 1. Por outro lado, o Avaliador 2 avaliou todos os conjuntos de imagens gerados pelo Modelo 01 como não estando de acordo com os *prompts* utilizados (Em desacordo).

- Modelo 02: As avaliadoras "Concordaram em parte"em quatro ocasiões que os conjuntos de imagens não saudáveis gerados pelo Modelo 02 estão alinhados com os prompts utilizados para gerá-los: três vezes pelo Avaliador 1 e uma vez pelo Avaliador 2 (conjuntos 2 e 4 da Figura 4.10f e Conjunto 1 da Figura 4.11f). Em um dos conjuntos avaliado, Conjunto 4 da Figura 4.10e, as avaliadoras foram unânimes em "Concorda em parte"de que o conjunto de imagens gerado está de acordo com o prompts utilizado para gerá-lo.

Nas figuras 4.12, 4.13, 4.14 e 4.15, são exibidas as avaliações agregadas de como os modelos se saíram na geração de imagens de raios-x do tórax. Cada barra nos gráficos representa a quantidade de vezes que os conjuntos de imagens, gerados por um modelo específico, receberam uma certa avaliação.

Por exemplo, uma barra pode mostrar quantas vezes um conjunto de imagens, gerado por um modelo treinado com 20 imagens por 20 épocas, foram classificadas -

quanto ao realismo das imagens - como "NÃO realista", "Realista"ou "MUITO realista". Da mesma forma, uma barra pode mostrar quantas vezes um conjunto de imagens, gerado por um modelo treinado em um *dataset* de 20 imagens por 20 épocas, foram classificadas em "Em desacordo", "Concorda em parte"ou "Concorda bastante", quanto à concordância do conjunto de imagens gerado com o *prompt* utilizado para a sua geração.

Para permitir uma melhor compreensão das informações apresentadas no gráfico, especificamos o que cada cor representa: vermelho para "NÃO realista" e "Em desacordo"; amarelo para "Realista" e "Concorda em parte"; verde para "MUITO realista" e "Concorda bastante".

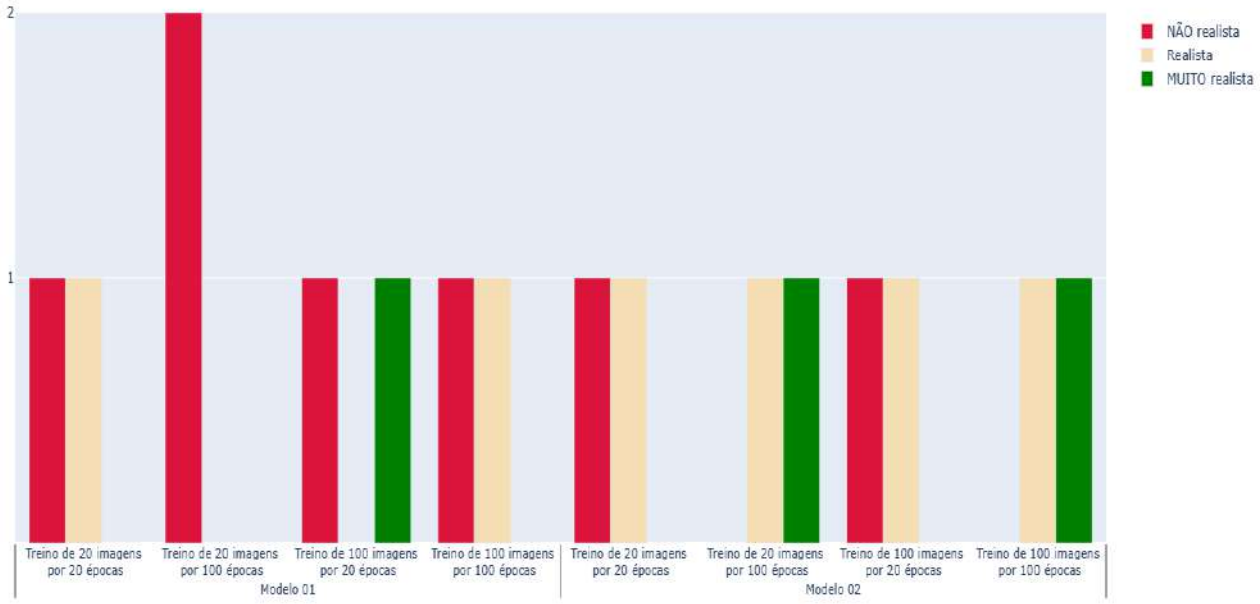


Figura 4.12: Avaliação feita por DUAS especialistas quanto ao desempenho dos modelos 1 e 2 na geração de QUATRO conjuntos de imagens de raios-x do tórax saudáveis, quanto ao realismo.

Com base nas avaliações realizadas nas duas fases, cujos resultados agregados são apresentados nos gráficos das Figuras 4.12, 4.13, 4.14 e 4.15, podemos afirmar que os modelos treinados utilizando o otimizador Adam8bit foram os que apresentaram melhor desempenho. Este desempenho superior foi observado tanto no quesito realismo das imagens geradas, quanto na conformidade das imagens geradas com os *prompts* utilizados para gerá-las.

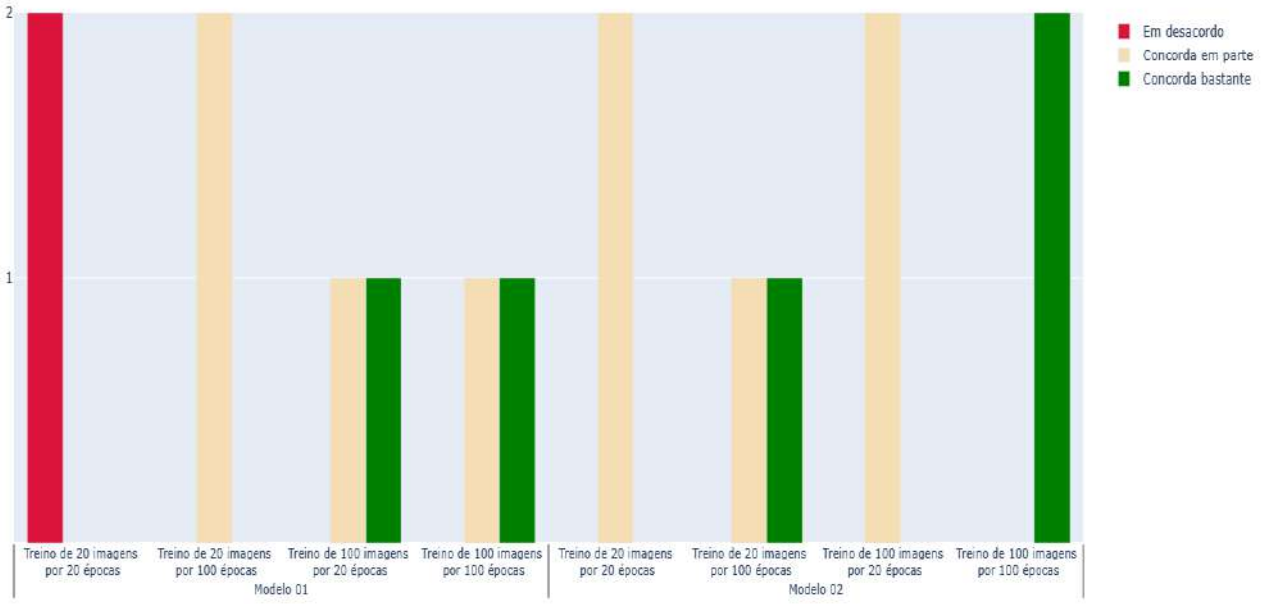


Figura 4.13: Avaliação feita por DUAS especialistas quanto ao desempenho dos modelos 1 e 2 na geração de QUATRO conjuntos de imagens de raios-x do tórax saudáveis, quanto ao *prompt*.

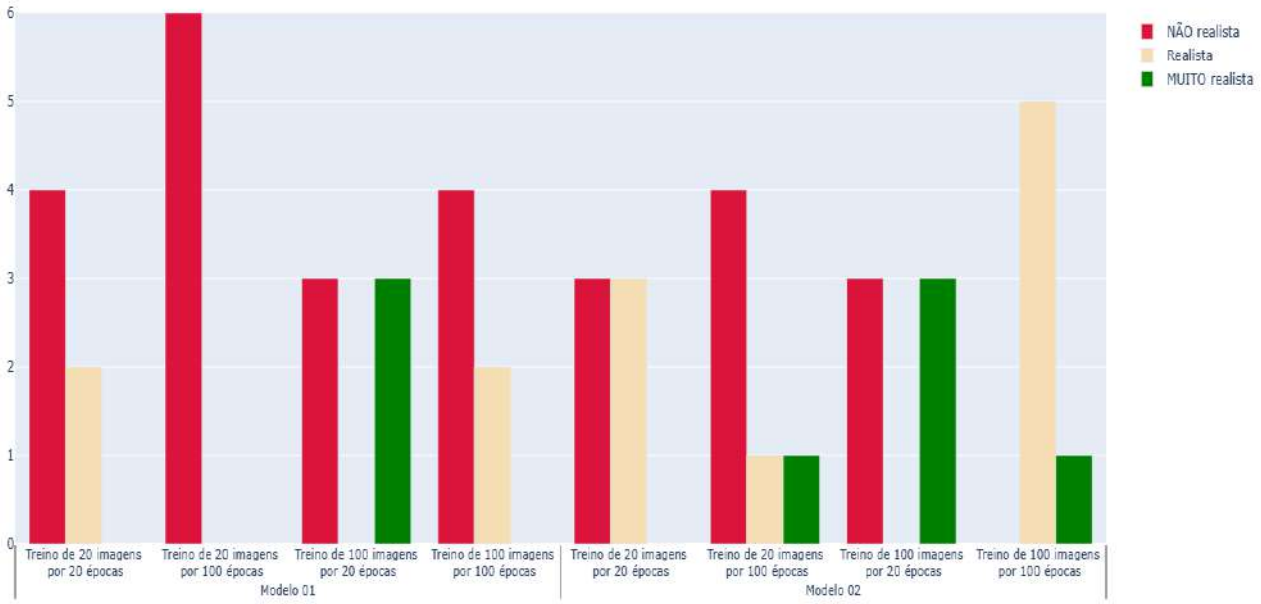


Figura 4.14: Avaliação feita por DUAS especialistas quanto ao desempenho dos modelos 1 e 2 na geração de QUATRO conjuntos de imagens de raios-x do tórax não saudáveis, quanto ao realismo.

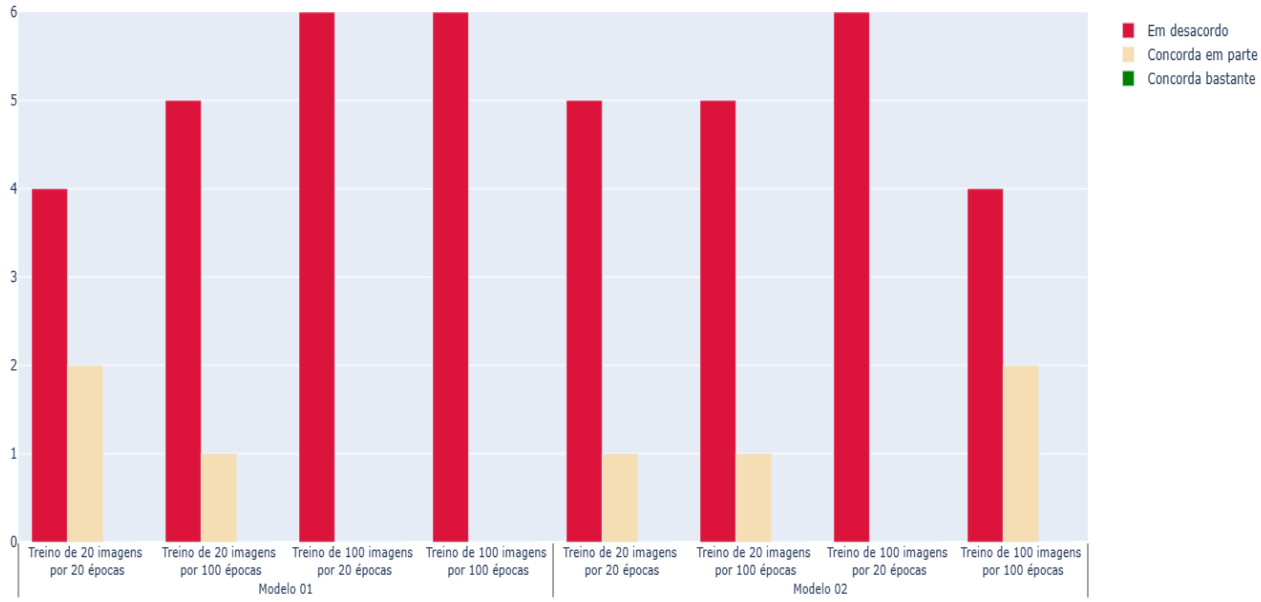


Figura 4.15: Avaliação feita por DUAS especialistas quanto ao desempenho dos modelos 1 e 2 na geração de QUATRO conjuntos de imagens de raios-x do tórax não saudáveis, quanto ao *prompt*.

4.3.3 Comparação de qualidade: imagens saudáveis vs. não saudáveis

Nesta seção, apresentamos uma comparação em termos da qualidade das imagens saudáveis e imagens não saudáveis geradas. Separamos as comparações em duas partes: na primeira parte, comparamos as qualidades das imagens saudáveis e não saudáveis geradas na primeira fase. Na segunda parte, comparamos as qualidades das imagens saudáveis e não saudáveis geradas na segunda fase.

Para as comparações em questão, adotamos os seguintes critérios: para os conjuntos das imagens gerados durante a primeira fase, classificamos como "BOM"aqueles conjuntos avaliados como sendo "RAZOAVELMENTE Realista", "BOM Realismo" e "MUITO Realista". Por outro lado, classificamos como "RUIM"aqueles conjuntos avaliados como "Totalmente NÃO Realista" e "POUCO Realista".

Para os conjuntos de imagens gerados na segunda fase, mantivemos um padrão semelhante ao da primeira. No entanto, expandimos os critérios de avaliação. Nesta etapa, consideramos como "BOM"os conjuntos que receberam avaliações de "Realista", "Muito realista", "Concorda em parte" e "Concorda bastante". Por outro lado, consideramos como "RUIM"aqueles conjuntos avaliados como "NÃO realista" e "Em desacordo"

Qualidade das imagens saudáveis e não saudáveis geradas na primeira fase

Quando comparamos a qualidade dos conjuntos das imagens geradas pelo *prompt* de geração de imagens saudáveis, classificados como BONS, e conjuntos das imagens geradas pelo *prompt* de geração de imagens não saudáveis, também classificados como BONS, temos o seguinte resultado:

Primeiramente, definimos a quantidade dos conjuntos das imagens testadas:

- Número Total dos Conjuntos de Imagens Saudáveis Avaliados: 4 (Totalmente NÃO Realista) + 0 (POUCO Realista) + 1 (RAZOAVELMENTE Realista) + 0 (BOM Realismo) + 1 (MUITO Realista) = 6
- Número Total dos Conjuntos de Imagens Não Saudáveis Avaliados: 4 (Totalmente NÃO Realista) + 1 (POUCO Realista) + 1 (RAZOAVELMENTE Realista) + 0 (BOM Realismo) + 0 (MUITO Realista) = 6

Em seguida, definimos o total dos conjuntos de imagens saudáveis e não saudáveis avaliados como BONS:

- Total dos conjuntos de avaliações de imagens saudáveis classificados como BONS: 1 (RAZOAVELMENTE Realista) + 0 (BOM Realismo) + 1 (MUITO Realista) = 2
- Total dos conjuntos de avaliações de imagens não saudáveis classificados como BONS: 1 (RAZOAVELMENTE Realista) + 0 (BOM Realismo) + 0 (MUITO Realista) = 1

Finalmente, calculamos a média dos conjuntos de imagens saudáveis e não saudáveis avaliados como BONS:

- Média dos conjuntos de imagens saudáveis avaliados como BONS: 2 (avaliações BOAS) / 6 (total de avaliações) = 0,33
- Média dos conjuntos de imagens não saudáveis avaliados como BONS: 1 (avaliações BOAS) / 6 (total de avaliações) = 0,17

Portanto, as imagens saudáveis (geradas com o prompt de geração de imagens saudáveis) receberam uma avaliação média mais alta (em termos de boa qualidade) em comparação com as imagens não saudáveis (geradas com o prompt de geração de imagens não saudáveis).

Qualidade das imagens saudáveis e não saudáveis geradas na segunda fase

Quando comparamos a qualidade dos conjuntos das imagens geradas pelo *prompt* de geração de imagens saudáveis classificados como BONS, e conjuntos das imagens geradas pelo *prompt* de geração de imagens não saudáveis, também classificados como BONS, temos o seguinte resultado:

Primeiramente, definimos a quantidade dos conjuntos das imagens testadas:

- Número Total dos Conjuntos de Imagens Saudáveis Avaliados: 7 (NÃO realista) + 6 (Realista) + 3 (MUITO realista) + 2 (Em desacordo) + 9 (Concorda em parte) + 5 (Concorda bastante) = 32
- Número Total dos Conjuntos de Imagens Não Saudáveis Avaliados: 28 (NÃO realista) + 14 (Realista) + 6 (MUITO realista) + 41 (Em desacordo) + 7 (Concorda em parte) + 0 (Concorda bastante) = 96

Em seguida, definimos o total dos conjuntos de imagens saudáveis e não saudáveis avaliados como BONS:

- Total dos conjuntos de avaliações de imagens saudáveis classificados como BONS: 6 (Realista) + 3 (MUITO realista) + 9 (Concorda em parte) + 5 (Concorda bastante) = 23
- Total dos conjuntos de avaliações de imagens não saudáveis classificados como BONS: 14 (Realista) + 6 (MUITO realista) + 7 (Concorda em parte) + 0 (Concorda bastante) = 27

Finalmente, calculamos a média dos conjuntos de imagens saudáveis e não saudáveis avaliados como BONS:

- Média dos Conjuntos de Imagens Saudáveis Avaliados como BONS: 23 (avaliações BOAS) / 32 (total de avaliações) = 0,72
- Média dos Conjuntos de Imagens Não Saudáveis Avaliados como BONS: 27 (avaliações BOAS) / 96 (total de avaliações) = 0,28

Portanto, as imagens saudáveis (geradas com o *prompt* de geração de imagens saudáveis) receberam uma avaliação média mais alta (em termos de boa qualidade) em comparação com as imagens não saudáveis (geradas com o *prompt* de geração de imagens não saudáveis).

Nas Tabelas 4.5 e 4.6 apresentamos as comparações gerais sobre a qualidade das imagens geradas tanto na primeira fase (Tabela 4.5) quanto na segunda fase (Tabela 4.6). Destacamos em vermelho a média dos conjuntos de imagens saudáveis avaliados como bons e a média dos conjuntos de imagens não saudáveis avaliados como bons. Em ambas as fases, observou-se que as imagens saudáveis, aquelas geradas a partir do *prompt* específico para essa categoria, receberam uma avaliação média superior em termos de qualidade. Esta avaliação foi comparativamente mais alta do que as imagens não saudáveis, que foram geradas a partir de *prompts* destinados a essa categoria. Especificamente, as imagens saudáveis obtiveram uma média de 0,3333 e 0,7188, enquanto as imagens não saudáveis obtiveram uma média de 0,1667 e 0,2813, respectivamente.

Tabela 4.5: Comparação de qualidade das imagens saudáveis e não saudáveis geradas na primeira fase.

Avaliações	Conjuntos de Imagens Saudáveis Avaliados	Conjuntos de Imagens Não Saudáveis Avaliados	Média dos Conjuntos de Imagens Saudáveis Avaliados como BONS	Média dos Conjuntos de Imagens Não Saudáveis Avaliados como BONS	Média dos Conjuntos de Imagens Saudáveis Avaliados como RUINS	Média dos Conjuntos de Imagens Não Saudáveis Avaliados como RUINS
Totalmente NÃO Realista	4	4	0,00 (0/6)	0,00 (0/6)	0,67 (4/6)	0,67 (4/6)
POUCO Realista	0	1	0,00 (0/6)	0,00 (0/6)	0,00 (0/6)	0,17 (1/6)
RAZOAVELMENTE Realista	1	1	0,17 (1/6)	0,17 (1/6)	0,00 (0/6)	0,00 (0/6)
BOM Realismo	0	0	0,00 (0/6)	0,00 (0/6)	0,00 (0/6)	0,00 (0/6)
MUITO Realista	1	0	0,17 (1/6)	0,00 (0/6)	0,00 (0/6)	0,00 (0/6)
Total	6	6	0,3333	0,1667	0,6667	0,8333

Tabela 4.6: Comparação de qualidade das imagens saudáveis e não saudáveis geradas na segunda fase.

Avaliações	Conjuntos de Imagens Saudáveis Avaliados	Conjuntos de Imagens Não Saudáveis Avaliados	Média dos Conjuntos de Imagens Saudáveis Avaliados como BONS	Média dos Conjuntos de Imagens Não Saudáveis Avaliados como BONS	Média dos Conjuntos de Imagens Saudáveis Avaliados como RUINS	Média dos Conjuntos de Imagens Não Saudáveis Avaliados como RUINS
NÃO realista	7	28	0,00 (0/32)	0,00 (0/96)	0,22 (7/32)	0,29 (28/96)
Realista	6	14	0,19 (6/32)	0,15 (14/96)	0,00 (0/32)	0,00 (0/96)
MUITO realista	3	6	0,09 (3/32)	0,06 (6/96)	0,00 (0/32)	0,00 (0/96)
Em desacordo	2	41	0,00 (0/32)	0,00 (0/96)	0,06 (2/32)	0,43 (41/96)
Concorda em parte	9	7	0,28 (9/32)	0,07 (7/96)	0,00 (0/32)	0,00 (0/96)
Concorda bastante	5	0	0,16 (5/32)	0,00 (0/96)	0,00 (0/32)	0,00 (0/96)
Total	32	96	0,7188	0,2813	0,2813	0,7188

Conjuntos mais bem avaliados

Para apresentar os conjuntos mais bem avaliados, adotamos um procedimento que se assemelha ao usado para comparar a qualidade dos conjuntos de imagens. Na primeira fase, o Conjunto 2 se destacou. Este foi gerado pelo modelo treinado com o otimizador Adam8bit, seguindo as configurações do Modelo 01, conforme apresentado na Tabela 3.6. Em seguida, veio o Conjunto 3, que também foi gerado por um modelo treinado com o otimizador Adam8bit, mas desta vez com as configurações do Modelo 02, também detalhadas na Tabela 3.6.

Para os conjuntos de imagens gerados na segunda fase, adotamos os seguintes critérios para determinar se eles foram avaliados como bons ou ruins: um conjunto é classificado como "BOM"se receber avaliações de "Realista", "Muito realista", "Concorda em parte"ou "Concorda bastante". Em contrapartida, se as avaliações forem "Não realista"ou "Em desacordo", o conjunto é categorizado como "RUIM".

Nas Tabelas 4.7 e 4.8 apresentamos os conjuntos de imagens saudáveis e não saudáveis mais bem avaliados quando é utilizado o *prompt* de geração de conjuntos de imagens saudáveis (Tabela 4.7) e quando são utilizados *prompts* de geração de conjuntos de imagens não saudáveis (Tabela 4.8).

Ao utilizar o *prompt* para a geração de conjuntos de imagens saudáveis, os conjuntos 3 e 4 destacam-se como os mais bem avaliados. O Conjunto 3 foi gerado por um modelo que passou por um treinamento de 20 épocas em um *dataset* composto por 100 imagens. Em contrapartida, o Conjunto 4 originou-se de um modelo treinado por um período mais extenso, 100 épocas, mas utilizando o mesmo *dataset* de 100 imagens.

Quando os *prompts* para a geração dos conjuntos de imagens não saudáveis são utilizados, os conjuntos que receberam as melhores avaliações são os conjuntos 1 e 4, conforme demonstrado na Tabela 4.8. Neste cenário, o Conjunto 1 foi produzido por um modelo treinado durante 20 épocas, mas em um *dataset* menor, de 20 imagens. Já o Conjunto 4, assim como no caso anterior, foi gerado por um modelo que passou por um treinamento de 100 épocas em um *dataset* de 100 imagens.

Tabela 4.7: Conjuntos de imagens saudáveis mais bem avaliados.

Avaliações	Conjunto 1 (Conjunto gerado pelo modelo treinado por 20 épocas em um *dataset* de 20 imagens)	Conjunto 2 (Conjunto gerado pelo modelo treinado por 20 épocas em um *dataset* de 100 imagens)	Conjunto 3 (Conjunto gerado pelo modelo treinado por 100 épocas em um *dataset* de 20 imagens)	Conjunto 4 (Conjunto gerado pelo modelo treinado por 100 épocas em um *dataset* de 100 imagens)
Total BOM	4	6	7	7
Total RUIM	4	2	1	1

Tabela 4.8: Conjuntos de imagens não saudáveis mais bem avaliados.

Avaliações	Conjunto 1 (Conjunto gerado pelo modelo treinado por 20 épocas em um *dataset* de 20 imagens)	Conjunto 2 (Conjunto gerado pelo modelo treinado por 20 épocas em um *dataset* de 100 imagens)	Conjunto 3 (Conjunto gerado pelo modelo treinado por 100 épocas em um *dataset* de 20 imagens)	Conjunto 4 (Conjunto gerado pelo modelo treinado por 100 épocas em um *dataset* de 100 imagens)
Total BOM	9	3	6	9
Total RUIM	15	21	18	15

Prompts mais bem avaliados

Na Tabela 4.9, apresentamos um comparativo do desempenho dos *prompts*. Dentre os *prompts* de geração de imagens não saudáveis utilizados, os que apresentaram melhor desempenho foram "*Human chest x-ray with right pleural effusion*" e "*Human chest x-ray with bilateral nodular infiltrate*". Esses resultados também estão detalhados na Tabela 4.9.

Tabela 4.9: *Prompts* de geração de conjuntos de imagens não saudáveis mais bem avaliados.

Avaliações	*Human chest x-ray with large area of cavitation in the right upper lobe*	*Human chest x-ray with bilateral nodular infiltrate*	*Human chest x-ray with right pleural effusion*
Total BOM	8	9	10
Total RUIM	24	23	22

Um aspecto adicional que merece ser destacado é o desempenho insatisfatório dos modelos na geração de imagens de raios-x que correspondam aos *prompts* utilizados (Figura 4.15). Esta situação pode estar potencialmente associada à ausência (no caso de termos compostos como "*pleural effusion*" e *nodular infiltrate*) ou à baixa ocorrência dos termos empregados nos *prompts*, nos conjuntos de dados utilizados para o treinamento dos modelos (os termos utilizados nos *prompts* de geração de imagens são apresentados nas Tabelas 3.4 e 3.5).

4.4 Requisitos Computacionais

O processamento de imagens exige uma alta capacidade computacional. Processar uma imagem é computacionalmente complexo, podendo envolver bilhões de cálculos para processar uma única imagem. Isso pode ser explicado devido ao grande volume de dados nas imagens e também devido à complexidade de algoritmos utilizados para processar essas imagens. Sendo assim, é recomendado utilizar hardwares especializados como as GPUs dado que estas foram projetadas para realizar operações complexas e paralelas e possuem capacidade para lidar com grande número de cálculos simultâneas exigidos no processamento de imagens. Portanto, neste trabalho foi utilizado a GPU da NVIDIA com as características abaixo listadas [76]:

- NVIDIA GeForce RTX 3050: é uma GPU da série NVIDIA GeForce RTX 30, cuja arquitetura é baseada na arquitetura NVIDIA Ampere. Ela é projetada para jogos e criação de conteúdo, oferecendo suporte a recursos como ray tracing e DLSS (Deep Learning Super Sampling). VRAM 8191: a GPU utilizada possui 8.191 megabytes (ou 8 gigabytes) de memória de vídeo, que é usada para armazenar dados necessários para renderizar imagens. É importante frisar também que quanto mais VRAM (Video Random-Access Memory) uma GPU tiver, mais dados ela poderá manipular sem diminuir a velocidade.
- Arch (8, 6) diz respeito à capacidade de computação da GPU. O número (8) indica a versão principal e o número (6) indica a versão secundária. Quanto maiores os números, mais avançada é a versão da GPU [44].
- Cores 20: a GPU utilizada possui 20 multiprocessadores, cada um contendo 128 núcleos cada. Portanto, o número total de núcleos na GPU é de 2.560 que podem executar instruções em paralelo [76, 98].

Nas Tabelas 4.10 e 4.11, os tempos de treinamento dos modelos são apresentados em horas. Na primeira fase, um conjunto de 30 imagens foi utilizado para os treinamentos. Todos os modelos passaram por cem épocas de treinamento, cujos tempos são detalhados na Tabela 4.10.

Na segunda fase, sete conjuntos de imagens foram utilizados para os treinamentos dos Modelos 01 e 02. Esses conjuntos variam em tamanho, começando com 5 imagens e aumentando sequencialmente até 100 imagens. Cada conjunto é um subconjunto do próximo, ou seja, o conjunto de 5 imagens está contido no conjunto de 10, que por sua vez está contido no conjunto de 20, e assim sucessivamente até o conjunto de 100 imagens. Ambos os modelos, Modelo 01 e Modelo 02, passaram por 5, 10, 20, 40, 60, 80 e 100 épocas de treinamento.

No entanto, na Tabela 4.11, decidimos destacar apenas os tempos de treinamento dos conjuntos de 20 e 100 imagens que passaram por treinamentos de 20 e 100 épocas, pois estes foram os que passaram pela avaliação dos especialistas. Todos os tempos de treinamento são apresentados em horas, para facilitar a compreensão.

Tabela 4.10: Tabela com as quantidades de imagens utilizadas para treinar os modelos na primeira fase e os tempos de treinamento dos modelos apresentados em horas.

Modelos	Quantidade de imagens utilizada no treinamento dos modelos	Tempo de treinamento para o modelo treinado por 100 épocas (horas)
Modelo 01	30 imagens	8,6269
Modelo 02		9,6014
Modelo 03		16,7983
Modelo 04		26,5325
Modelo 05		38,1503

Tabela 4.11: Tabela com as quantidades de imagens utilizadas para treinar os modelos na segunda fase e os tempos de treinamento dos modelos apresentados em horas.

Modelos	Quantidade de imagens utilizada no treinamento	Tempo de treinamento para o modelo treinado por 20 épocas (horas)	Tempo de treinamento para o modelo treinado por 100 épocas (horas)
Modelo 01	20 imagens	1,1236	6,6025
	100 imagens	3,2617	17,0308
Modelo 02	20 imagens	0,8214	4,6292
	100 imagens	3,5670	18,75111

5. CONSIDERAÇÕES FINAIS

Neste trabalho, apresentamos a aplicação de um método de geração de imagens de raios-x do tórax utilizando o *Latent Diffusion Model*. Nossa abordagem permite que o usuário realize *fine-tuning* no *Latent Diffusion Model* e treine os modelos utilizando ferramentas como o LoRA. Além disso, nosso método possibilita gerar imagens de raios-x do tórax aplicando os modelos treinados na ferramenta AUTOMATIC1111. Essa ferramenta possui uma interface que permite selecionar os modelos treinados, inserir *prompts* de texto e ajustar parâmetros adicionais para gerar características específicas desejadas nas imagens.

Análises de desempenho dos modelos na geração de imagens de raios-x também foram conduzidas por médicos. Essas análises de desempenho ocorreram em duas etapas: na primeira, um único médico conduziu as análises, e na segunda, duas médicas realizaram as avaliações. Em ambas as etapas, os modelos treinados com o otimizador Adam8bit demonstraram melhor desempenho na geração de imagens de raios-x realistas e na produção de imagens que atendem aos *prompts*.

É importante destacar que, na primeira fase do nosso trabalho, utilizamos quatro tipos de otimizadores diferentes (AdamW8bit, Adafactor, DAdaptSGD e Prodigy) e um conjunto de dados composto por apenas 30 imagens (50% de imagens saudáveis e 50% de imagens não saudáveis) para treinar os modelos. Após avaliações médicas, constatou-se que os modelos treinados com o otimizador adam8bit apresentaram o melhor desempenho na geração de imagens realistas. Com base nesses resultados, avançamos para a segunda fase dos treinamentos.

Na segunda fase, expandimos nosso trabalho para incluir sete conjuntos de dados de tamanhos diferentes, cada um contendo 5, 10, 20, 40, 60, 80 e 100 imagens, respectivamente (com 50% de imagens saudáveis e 50% de imagens não saudáveis em cada conjunto). Geramos imagens a partir desses conjuntos, mas devido ao número insuficiente de especialistas, decidimos submeter à avaliação apenas as imagens geradas pelos modelos treinados nos conjuntos de dados de 20 e 100 imagens.

Após a avaliação, verificou-se que o Modelo 02, treinado com um conjunto de dados contendo 100 imagens, obteve o melhor desempenho em comparação com os demais modelos. Esse resultado reforça a eficácia do otimizador adam8bit no treinamento de modelos para a geração de imagens realistas.

5.1 Limitações

Ao longo do desenvolvimento deste trabalho, deparamos com três desafios principais. O primeiro foi a dificuldade em localizar um conjunto de dados que contivesse imagens de raios-x do tórax, juntamente com seus respectivos textos descritivos ou relatórios clínicos. Este desafio surge porque esses conjuntos de dados são altamente especializados e podem envolver questões delicadas, como privacidade e consentimento do paciente, para acessá-los. Além disso, a qualidade e a consistência dos relatórios clínicos podem variar, comprometendo a utilidade do conjunto de dados para o treinamento de modelos.

O segundo desafio foi a necessidade de máquinas com recursos computacionais robustos para processar essas imagens. A disponibilidade limitada desses recursos dificultou a busca por uma solução adequada, restringindo a complexidade dos modelos que poderiam ser treinados e a velocidade com que os modelos poderiam ser desenvolvidos e testados.

Por fim, o terceiro desafio com o qual deparamos o fato de não termos médicos suficientes para validar os testes. A título de exemplo, cita-se o fato de não conseguirmos obter a assistência de um radiologista especialista em tuberculose. Esta situação ressalta a importância de se ter profissionais médicos especialistas disponíveis para validar e interpretar os resultados, garantindo a precisão e a eficácia dos modelos treinados ou desenvolvidos.

5.2 Trabalhos Futuros

Para os trabalhos futuros, o nosso objetivo é superar os desafios encontrados neste estudo buscando, primeiramente, expandir nossas fontes de dados para obter um conjunto de dados mais robusto e diversificado. Isso inclui imagens de raios-x do tórax e seus respectivos textos descritivos ou relatórios clínicos. Para isso, buscaremos parcerias com instituições que possam fornecer o acesso a esses dados.

Adicionalmente, planejamos explorar soluções alternativas que possam auxiliar no processamento de imagens. Isso pode envolver o uso de algoritmos mais eficientes ou trabalhar em colaboração com instituições que possam fornecer acesso a recursos computacionais mais potentes.

Finalmente, planejamos expandir a aplicabilidade de nosso trabalho para incluir não apenas a área de saúde, mas também a área de educação. Para isso, estamos considerando o desenvolvimento de aplicações que sejam úteis tanto no contexto da saúde quanto no da educação. Pensamos numa aplicação onde os professores possam utilizar

nosso método para gerar exemplos personalizados do que gostariam de trabalhar com os alunos, enriquecendo assim o conteúdo didático apresentado aos alunos. Acreditamos que essa abordagem pode aprimorar o processo de ensino e aprendizagem, proporcionando uma experiência educacional interativa. Além disso, acreditamos também que essa abordagem irá oferecer uma nova dimensão ao processo de ensino, tornando-o um processo mais adaptável e relevante.

REFERÊNCIAS BIBLIOGRÁFICAS

[1] Aerts, S.; Lambrechts, D.; Maity, S.; Van Loo, P.; Coessens, B.; De Smet, F.; Tranchevent, L.-C.; De Moor, B.; Marynen, P.; Hassan, B.; et al.. “Gene prioritization through genomic data fusion”, *Nature biotechnology*, vol. 24–5, 2006, pp. 537–544.

[2] Ali, H.; Murad, S.; Shah, Z. “Spot the fake lungs: Generating synthetic medical images using neural diffusion models”. In: Irish Conference on Artificial Intelligence and Cognitive Science, 2022, pp. 32–39.

[3] Amann, J.; Blasimme, A.; Vayena, E.; Frey, D.; Madai, V. I.; Consortium, P. “Explainability for artificial intelligence in healthcare: a multidisciplinary perspective”, *BMC medical informatics and decision making*, vol. 20, 2020, pp. 1–9.

[4] Antoniou, A.; Storkey, A.; Edwards, H. “Data augmentation generative adversarial networks”, *arXiv preprint arXiv:1711.04340*, 2017.

[5] AUTOMATIC1111. “Stable diffusion web ui”. Capturado em: https://github.com/AUTOMATIC1111/stable-diffusion-webui, Jun 2023.

[6] Azad, B.; Azad, R.; Eskandari, S.; Bozorgpour, A.; Kazerouni, A.; Rekik, I.; Merhof, D. “Foundational models in medical imaging: A comprehensive survey and future vision”. 2310.18689, 2023.

[7] Beck, J. T.; Rammage, M.; Jackson, G. P.; Preininger, A. M.; Dankwa-Mullan, I.; Roebuck, M. C.; Torres, A.; Holtzen, H.; Coverdill, S. E.; Williamson, M. P.; et al.. “Artificial intelligence tool for optimizing eligibility screening for clinical trials in a large community cancer center”, *JCO clinical cancer informatics*, vol. 4, 2020, pp. 50–59.

[8] Benam, K. H.; Gilchrist, S.; Kleensang, A.; Satz, A. B.; Willett, C.; Zhang, Q. “Exploring new technologies in biomedical research”, *Drug discovery today*, vol. 24–6, 2019, pp. 1242–1247.

[9] bmaltais. “Kohya’s gui”. Capturado em: https://github.com/bmaltais/kohya_ss, Jun 2023.

[10] Bommasani, R.; Hudson, D. A.; Adeli, E.; Altman, R.; Arora, S.; von Arx, S.; Bernstein, M. S.; Bohg, J.; Bosselut, A.; Brunskill, E.; et al.. “On the opportunities and risks of foundation models”, *arXiv preprint arXiv:2108.07258*, vol. Nothing, 2021.

[11] Brown, T.; Mann, B.; Ryder, N.; Subbiah, M.; Kaplan, J. D.; Dhariwal, P.; Neelakantan, A.; Shyam, P.; Sastry, G.; Askell, A.; et al.. “Language models are few-shot learners”, *Advances in neural information processing systems*, vol. 33, 2020, pp. 1877–1901.

[12] Cetin, I.; Stephens, M.; Camara, O.; Ballester, M. A. G. “Attri-vae: Attribute-based interpretable representations of medical images with variational autoencoders”, *Computerized Medical Imaging and Graphics*, vol. 104, 2023, pp. 102158.

[13] Chaitanya, K.; Erdil, E.; Karani, N.; Konukoglu, E. “Contrastive learning of global and local features for medical image segmentation with limited annotations”, *Advances in neural information processing systems*, vol. 33, 2020, pp. 12546–12558.

[14] Challen, R.; Denny, J.; Pitt, M.; Gompels, L.; Edwards, T.; Tsaneva-Atanasova, K. “Artificial intelligence, bias and clinical safety”, *BMJ quality & safety*, 2019.

[15] Chambon, P.; Bluethgen, C.; Delbrouck, J.-B.; Van der Sluijs, R.; Połacin, M.; Chaves, J. M. Z.; Abraham, T. M.; Purohit, S.; Langlotz, C. P.; Chaudhari, A. “Roentgen: vision-language foundation model for chest x-ray generation”, *arXiv preprint arXiv:2211.12737*, 2022.

[16] Chamikara, M. A. P.; Bertok, P.; Khalil, I.; Liu, D.; Camtepe, S. “Privacy preserving distributed machine learning with federated learning”, *Computer Communications*, vol. 171, 2021, pp. 112–125.

[17] Chatterjee, S.; Maity, S.; Bhattacharjee, M.; Banerjee, S.; Das, A. K.; Ding, W. “Variational autoencoder based imbalanced covid-19 detection using chest x-ray images”, *New Generation Computing*, vol. 41–1, 2023, pp. 25–60.

[18] Chen, I. Y.; Joshi, S.; Ghassemi, M. “Treating health disparities with artificial intelligence”, *Nature medicine*, vol. 26–1, 2020, pp. 16–17.

[19] Chen, I. Y.; Szolovits, P.; Ghassemi, M. “Can ai help reduce disparities in general medical and mental health care?”, *AMA journal of ethics*, vol. 21–2, 2019, pp. 167–179.

[20] Chen, W.; Hu, H.; Saharia, C.; Cohen, W. W. “Re-imagen: Retrieval-augmented text-to-image generator”, *arXiv preprint arXiv:2209.14491*, 2022.

[21] Creswell, A.; White, T.; Dumoulin, V.; Arulkumaran, K.; Sengupta, B.; Bharath, A. A. “Generative adversarial networks: An overview”, *IEEE signal processing magazine*, vol. 35–1, 2018, pp. 53–65.

[22] Cuenca, P.; Paul, S. “Using lora for efficient stable diffusion fine-tuning”. Capturado em: https://huggingface.co/blog/lora, Oct 2023.

[23] Daniel, J. E.; Brink, W.; Eloff, R.; Copley, C. “Towards automating healthcare question answering in a noisy multilingual low-resource setting”. In: Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics, 2019, pp. 948–953.

[24] Davenport, T.; Kalakota, R. “The potential for artificial intelligence in healthcare”, *Future healthcare journal*, vol. 6–2, 2019, pp. 94.

[25] Defazio, A.; Mishchenko, K. “Learning-rate-free learning by d-adaptation”. In: International Conference on Machine Learning, 2023, pp. 7449–7479.

[26] Dettmers, T.; Lewis, M.; Shleifer, S.; Zettlemoyer, L. “8-bit optimizers via block-wise quantization”, *arXiv preprint arXiv:2110.02861*, 2021.

[27] Dhariwal, P.; Nichol, A. “Diffusion models beat gans on image synthesis”, *Advances in neural information processing systems*, vol. 34, 2021, pp. 8780–8794.

[28] Dou, H.; Chen, C.; Hu, X.; Xuan, Z.; Hu, Z.; Peng, S. “Pca-srgan: Incremental orthogonal projection discrimination for face super-resolution”. In: Proceedings of the 28th ACM International Conference on Multimedia, 2020, pp. 1891–1899.

[29] Esteva, A.; Robicquet, A.; Ramsundar, B.; Kuleshov, V.; DePristo, M.; Chou, K.; Cui, C.; Corrado, G.; Thrun, S.; Dean, J. “A guide to deep learning in healthcare”, *Nature medicine*, vol. 25–1, 2019, pp. 24–29.

[30] Face, H. “8-bit optimizers”. Capturado em: https://huggingface.co/docs/bitsandbytes/main/en/optimizers, Jun 2023.

[31] Face, H. “Stable diffusion v1-5 model card”. Capturado em: https://huggingface.co/runwayml/stable-diffusion-v1-5, Jun 2023.

[32] Gao, P.; Geng, S.; Zhang, R.; Ma, T.; Fang, R.; Zhang, Y.; Li, H.; Qiao, Y. “Clip-adapter: Better vision-language models with feature adapters”, *International Journal of Computer Vision*, 2023, pp. 1–15.

[33] Goncalves, A.; Ray, P.; Soper, B.; Stevens, J.; Coyle, L.; Sales, A. P. “Generation and evaluation of synthetic patient data”, *BMC medical research methodology*, vol. 20–1, 2020, pp. 1–40.

[34] Goodfellow, I.; Pouget-Abadie, J.; Mirza, M.; Xu, B.; Warde-Farley, D.; Ozair, S.; Courville, A.; Bengio, Y. “Generative adversarial networks”, *Communications of the ACM*, vol. 63–11, 2020, pp. 139–144.

[35] Guo, Y.; Yang, C.; Rao, A.; Wang, Y.; Qiao, Y.; Lin, D.; Dai, B. “Animatediff: Animate your personalized text-to-image diffusion models without specific tuning”, *arXiv preprint arXiv:2307.04725*, 2023.

[36] Guu, K.; Lee, K.; Tung, Z.; Pasupat, P.; Chang, M. “Retrieval augmented language model pre-training”. In: International conference on machine learning, 2020, pp. 3929–3938.

[37] Han, K.; Xiong, Y.; You, C.; Khosravi, P.; Sun, S.; Yan, X.; Duncan, J. S.; Xie, X. “Medgen3d: A deep generative framework for paired 3d image and mask generation”. In: International Conference on Medical Image Computing and Computer-Assisted Intervention, 2023, pp. 759–769.

[38] Harrer, S.; Shah, P.; Antony, B.; Hu, J. “Artificial intelligence for clinical trial design”, *Trends in pharmacological sciences*, vol. 40–8, 2019, pp. 577–591.

[39] He, K.; Zhang, X.; Ren, S.; Sun, J. “Deep residual learning for image recognition”. In: Proceedings of the IEEE conference on computer vision and pattern recognition, 2016, pp. 770–778.

[40] Ho, J.; Jain, A.; Abbeel, P. “Denoising diffusion probabilistic models”, *Advances in neural information processing systems*, vol. 33, 2020, pp. 6840–6851.

[41] Holzinger, A.; Langs, G.; Denk, H.; Zatloukal, K.; Müller, H. “Causability and explainability of artificial intelligence in medicine”, *Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery*, vol. 9–4, 2019, pp. e1312.

[42] Houlsby, N.; Giurgiu, A.; Jastrzebski, S.; Morrone, B.; De Laroussilhe, Q.; Gesmundo, A.; Attariyan, M.; Gelly, S. “Parameter-efficient transfer learning for nlp”. In: International Conference on Machine Learning, 2019, pp. 2790–2799.

[43] Hu, E. J.; Shen, Y.; Wallis, P.; Allen-Zhu, Z.; Li, Y.; Wang, S.; Wang, L.; Chen, W. “Lora: Low-rank adaptation of large language models”, *arXiv preprint arXiv:2106.09685*, 2021.

[44] IBM. “Arch”. Capturado em: https://www.ibm.com/docs/en/cobol-zos/6.3?topic=v6-arch, Oct 2023.

[45] Jaeger, S.; Candemir, S.; Antani, S.; Wáng, Y.-X. J.; Lu, P.-X.; Thoma, G. “Two public chest x-ray datasets for computer-aided screening of pulmonary diseases”, *Quantitative imaging in medicine and surgery*, vol. 4–6, 2014, pp. 475.

[46] Jaroch, K.; Jaroch, A.; Bojko, B. “Cell cultures in drug discovery and development: The need of reliable in vitro-in vivo extrapolation for pharmacodynamics and pharmacokinetics assessment”, *Journal of Pharmaceutical and Biomedical Analysis*, vol. 147, 2018, pp. 297–312.

[47] Ji, Y.; Zhou, Z.; Liu, H.; Davuluri, R. V. “Dnabert: pre-trained bidirectional encoder representations from transformers model for dna-language in genome”, *Bioinformatics*, vol. 37–15, 2021, pp. 2112–2120.

[48] Jiang, L.; Mao, Y.; Wang, X.; Chen, X.; Li, C. “Cola-diff: Conditional latent diffusion model for multi-modal mri synthesis”. In: International Conference on Medical Image Computing and Computer-Assisted Intervention, 2023, pp. 398–408.

[49] Jin, Q.; Yuan, Z.; Xiong, G.; Yu, Q.; Ying, H.; Tan, C.; Chen, M.; Huang, S.; Liu, X.; Yu, S. “Biomedical question answering: a survey of approaches and challenges”, *ACM Computing Surveys (CSUR)*, vol. 55–2, 2022, pp. 1–36.

[50] Jumper, J.; Evans, R.; Pritzel, A.; Green, T.; Figurnov, M.; Tunyasuvunakool, K.; Ronneberger, O.; Bates, R.; Žídek, A.; Bridgland, A.; et al.. “Alphafold 2”, *Fourteenth Critical Assessment of Techniques for Protein Structure Prediction; DeepMind: London, UK*, 2020.

[51] Kadurin, A.; Nikolenko, S.; Khrabrov, K.; Aliper, A.; Zhavoronkov, A. “drugan: An advanced generative adversarial autoencoder model for de novo generation of new molecules with desired molecular properties in silico”, *Molecular Pharmaceutics*, vol. 14–9, 2017, pp. 3098–3104, pMID: 28703000, https://doi.org/10.1021/acs.molpharmaceut.7b00346.

[52] Kaushal, A.; Altman, R.; Langlotz, C. “Geographic distribution of us cohorts used to train deep learning algorithms”, *Jama*, vol. 324–12, 2020, pp. 1212–1213.

[53] Kazerouni, A.; Aghdam, E. K.; Heidari, M.; Azad, R.; Fayyaz, M.; Hacihaliloglu, I.; Merhof, D. “Diffusion models in medical imaging: A comprehensive survey”, *Medical Image Analysis*, 2023, pp. 102846.

[54] Kenton, J. D. M.-W. C.; Toutanova, L. K. “Bert: Pre-training of deep bidirectional transformers for language understanding”. In: Proceedings of naacL-HLT, 2019, pp. 2.

[55] Khader, F.; Müller-Franzes, G.; Tayebi Arasteh, S.; Han, T.; Haarburger, C.; Schulze-Hagen, M.; Schad, P.; Engelhardt, S.; Baeßler, B.; Foersch, S.; et al.. “Denoising diffusion probabilistic models for 3d medical image generation”, *Scientific Reports*, vol. 13–1, 2023, pp. 7303.

[56] Kingma, D. P.; Ba, J. “Adam: A method for stochastic optimization”. 1412.6980, 2017.

[57] Kirch, D. G.; Petelle, K. “Addressing the physician shortage: the peril of ignoring demography”, *Jama*, vol. 317–19, 2017, pp. 1947–1948.

[58] Klasnja, P.; Pratt, W. “Healthcare in the pocket: mapping the space of mobile-phone health interventions”, *Journal of biomedical informatics*, vol. 45–1, 2012, pp. 184–198.

[59] Kong, J.; Cooper, L. A.; Wang, F.; Gutman, D. A.; Gao, J.; Chisolm, C.; Sharma, A.; Pan, T.; Van Meir, E. G.; Kurc, T. M.; et al.. “Integrative, multimodal analysis of glioblastoma using tcga molecular data, pathology images, and clinical outcomes”, *IEEE Transactions on Biomedical Engineering*, vol. 58–12, 2011, pp. 3469–3474.

[60] Kourou, K.; Exarchos, T. P.; Exarchos, K. P.; Karamouzis, M. V.; Fotiadis, D. I. “Machine learning applications in cancer prognosis and prediction”, *Computational and structural biotechnology journal*, vol. 13, 2015, pp. 8–17.

[61] Krumholz, H. M. “Big data and new knowledge in medicine: the thinking, training, and tools needed for a learning health system”, *Health Affairs*, vol. 33–7, 2014, pp. 1163–1170.

[62] Lanckriet, G. R.; De Bie, T.; Cristianini, N.; Jordan, M. I.; Noble, W. S. “A statistical framework for genomic data fusion”, *Bioinformatics*, vol. 20–16, 2004, pp. 2626–2635.

[63] Lavertu, A.; Altman, R. B. “Redmed: Extending drug lexicons for social media applications”, *Journal of biomedical informatics*, vol. 99, 2019, pp. 103307.

[64] Lee, J.; Yoon, W.; Kim, S.; Kim, D.; Kim, S.; So, C. H.; Kang, J. “Biobert: a pre-trained biomedical language representation model for biomedical text mining”, *Bioinformatics*, vol. 36–4, 2020, pp. 1234–1240.

[65] Li, I.; Yasunaga, M.; Nuzumlalı, M. Y.; Caraballo, C.; Mahajan, S.; Krumholz, H.; Radev, D. “A neural topic-attention model for medical term abbreviation disambiguation”, *arXiv preprint arXiv:1910.14076*, 2019.

[66] Lin, Z.; Yu, S.; Kuang, Z.; Pathak, D.; Ramanan, D. “Multimodality helps unimodality: Cross-modal few-shot learning with multimodal models”. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2023, pp. 19325–19337.

[67] Linardatos, P.; Papastefanopoulos, V.; Kotsiantis, S. “Explainable ai: A review of machine learning interpretability methods”, *Entropy*, vol. 23–1, 2020, pp. 18.

[68] Liu, F.; Zhang, T.; Dai, W.; Cai, W.; Zhou, X.; Chen, D. “Few-shot adaptation of multi-modal foundation models: A survey”, *arXiv preprint arXiv:2401.01736*, vol. Nothing, 2024.

[69] Ma, J.; Fong, S. H.; Luo, Y.; Bakkenist, C. J.; Shen, J. P.; Mourragui, S.; Wessels, L. F.; Hafner, M.; Sharan, R.; Peng, J.; et al.. “Few-shot learning creates predictive models of drug response that translate from high-throughput screens to individual patients”, *Nature Cancer*, vol. 2–2, 2021, pp. 233–244.

[70] Madani, A.; Moradi, M.; Karargyris, A.; Syeda-Mahmood, T. “Chest x-ray generation and data augmentation for cardiovascular abnormality classification”. In: Medical imaging 2018: Image processing, 2018, pp. 415–420.

[71] Martinez-Martin, N.; Luo, Z.; Kaushal, A.; Adeli, E.; Haque, A.; Kelly, S. S.; Wieten, S.; Cho, M. K.; Magnus, D.; Fei-Fei, L.; et al.. “Ethical issues in using ambient intelligence in health-care settings”, *The lancet digital health*, vol. 3–2, 2021, pp. e115–e123.

[72] Mishchenko, K.; Defazio, A. “Prodigy: An expeditiously adaptive parameter-free learner”, *arXiv preprint arXiv:2306.06101*, 2023.

[73] Motamed, S.; Rogalla, P.; Khalvati, F. “Data augmentation using generative adversarial networks (gans) for gan-based detection of pneumonia and covid-19 in chest x-ray images”, *Informatics in Medicine Unlocked*, vol. 27, 2021, pp. 100779.

[74] Mozannar, H.; Sontag, D. “Consistent estimators for learning to defer to an expert”. In: International Conference on Machine Learning, 2020, pp. 7076–7087.

[75] Nie, A.; Zehnder, A.; Page, R. L.; Zhang, Y.; Pineda, A. L.; Rivas, M. A.; Bustamante, C. D.; Zou, J. “Deeptag: inferring diagnoses from veterinary clinical notes”, *NPJ digital medicine*, vol. 1–1, 2018, pp. 60.

[76] NVIDIA. “Geforce rtx 3050”. Capturado em: https://www.nvidia.com/pt-br/geforce/graphics-cards/30-series/rtx-3050/, Oct 2023.

[77] Oussidi, A.; Elhassouny, A. “Deep generative models: Survey”. In: 2018 International conference on intelligent systems and computer vision (ISCV), 2018, pp. 1–8.

[78] Packhäuser, K.; Folle, L.; Thamm, F.; Maier, A. “Generation of anonymous chest radiographs using latent diffusion models for training thoracic abnormality classification systems”. In: 2023 IEEE 20th International Symposium on Biomedical Imaging (ISBI), 2023, pp. 1–5.

[79] Pan, S. J.; Yang, Q. “A survey on transfer learning”, *IEEE Transactions on knowledge and data engineering*, vol. 22–10, 2009, pp. 1345–1359.

[80] Peters, M. E.; Neumann, M.; Iyyer, M.; Gardner, M.; Clark, C.; Lee, K.; Zettlemoyer, L. “Deep contextualized word representations. naacl-hlt”, *arXiv*, vol. Nothing–Nothing, 2018.

[81] Petroni, F.; Rocktäschel, T.; Lewis, P.; Bakhtin, A.; Wu, Y.; Miller, A. H.; Riedel, S. “Language models as knowledge bases?”, *arXiv preprint arXiv:1909.01066*, vol. Nothing, 2019.

[82] Pinaya, W. H.; Tudosiu, P.-D.; Dafflon, J.; Da Costa, P. F.; Fernandez, V.; Nachev, P.; Ourselin, S.; Cardoso, M. J. “Brain imaging generation with latent diffusion models”. In: MICCAI Workshop on Deep Generative Models, 2022, pp. 117–126.

[83] Radford, A.; Kim, J. W.; Hallacy, C.; Ramesh, A.; Goh, G.; Agarwal, S.; Sastry, G.; Askell, A.; Mishkin, P.; Clark, J.; et al.. “Learning transferable visual models from natural language supervision”. In: International conference on machine learning, 2021, pp. 8748–8763.

[84] Rai, R.; Tiwari, M. K.; Ivanov, D.; Dolgui, A. “Machine learning in manufacturing and industry 4.0 applications”, 2021.

[85] Ramesh, A.; Pavlov, M.; Goh, G.; Gray, S.; Voss, C.; Radford, A.; Chen, M.; Sutskever, I. “Zero-shot text-to-image generation”. In: International Conference on Machine Learning, 2021, pp. 8821–8831.

[86] Reed, S.; Akata, Z.; Yan, X.; Logeswaran, L.; Schiele, B.; Lee, H. “Generative adversarial text to image synthesis”. In: International conference on machine learning, 2016, pp. 1060–1069.

[87] Ribeiro, R. T.; Marinho, R. T.; Sanches, J. M. “Classification and staging of chronic liver disease from multimodal data”, *IEEE transactions on biomedical engineering*, vol. 60–5, 2012, pp. 1336–1344.

[88] Roberts, A.; Raffel, C.; Shazeer, N. “How much knowledge can you pack into the parameters of a language model?”, *arXiv preprint arXiv:2002.08910*, 2020.

[89] Rombach, R.; Blattmann, A.; Lorenz, D.; Esser, P.; Ommer, B. “High-resolution image synthesis with latent diffusion models”. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition, 2022, pp. 10684–10695.

[90] Ruiz, C.; Zitnik, M.; Leskovec, J. “Identification of disease treatment mechanisms through the multiscale interactome”, *Nature communications*, vol. 12–1, 2021, pp. 1796.

[91] Ruiz, N.; Li, Y.; Jampani, V.; Pritch, Y.; Rubinstein, M.; Aberman, K. “Dreambooth: Fine tuning text-to-image diffusion models for subject-driven generation”. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2023, pp. 22500–22510.

[92] Shazeer, N.; Stern, M. “Adafactor: Adaptive learning rates with sublinear memory cost”. 1804.04235, 2018.

[93] Singh, A.; Ogunfunmi, T. “An overview of variational autoencoders for source separation, finance, and bio-signal applications”, *Entropy*, vol. 24–1, 2021, pp. 55.

[94] Sinsky, C.; Colligan, L.; Li, L.; Prgomet, M.; Reynolds, S.; Goeders, L.; Westbrook, J.; Tutty, M.; Blike, G. “Allocation of physician time in ambulatory practice: a time and

motion study in 4 specialties", *Annals of internal medicine*, vol. 165–11, 2016, pp. 753–760.

[95] Snell, J.; Swersky, K.; Zemel, R. "Prototypical networks for few-shot learning", *Advances in neural information processing systems*, vol. 30, 2017.

[96] Sundaram, S.; Hulkund, N. "Gan-based data augmentation for chest x-ray classification". 2107.02970, 2021.

[97] Taylor, M. E.; Stone, P. "Transfer learning for reinforcement learning domains: A survey.", *Journal of Machine Learning Research*, vol. 10–7, 2009.

[98] techpowerup. "Nvidia geforce rtx 3050 8 gb". Capturado em: https://www.techpowerup.com/gpu-specs/geforce-rtx-3050-8-gb.c3858, Oct 2023.

[99] Thanh-Tung, H.; Tran, T. "Catastrophic forgetting and mode collapse in gans". In: 2020 international joint conference on neural networks (ijcnn), 2020, pp. 1–10.

[100] Thapa, C.; Camtepe, S. "Precision health data: Requirements, challenges and existing techniques for data security and privacy", *Computers in biology and medicine*, vol. 129, 2021, pp. 104130.

[101] Tudosiu, P.-D.; Pinaya, W. H. L.; Graham, M. S.; Borges, P.; Fernandez, V.; Yang, D.; Appleyard, J.; Novati, G.; Mehra, D.; Vella, M.; Nachev, P.; Ourselin, S.; Cardoso, J. "Morphology-preserving autoregressive 3d generative modelling of the brain". In: Simulation and Synthesis in Medical Imaging, Zhao, C.; Svoboda, D.; Wolterink, J. M.; Escobar, M. (Editores), 2022, pp. 66–78.

[102] Wang, X.; Xing, E. P.; Schaid, D. J. "Kernel methods for large-scale genomic data analysis", *Briefings in bioinformatics*, vol. 16–2, 2015, pp. 183–192.

[103] Wang, Y.; Li, J.; Naumann, T.; Xiong, C.; Cheng, H.; Tinn, R.; Wong, C.; Usuyama, N.; Rogahn, R.; Shen, Z.; et al.. "Domain-specific pretraining for vertical search: Case study on biomedical literature". In: Proceedings of the 27th ACM SIGKDD Conference on Knowledge Discovery & Data Mining, 2021, pp. 3717–3725.

[104] Wang, Z.; Wu, Z.; Agarwal, D.; Sun, J. "Medclip: Contrastive learning from unpaired medical images and text". 2210.10163, 2022.

[105] Wiens, J.; Saria, S.; Sendak, M.; Ghassemi, M.; Liu, V. X.; Doshi-Velez, F.; Jung, K.; Heller, K.; Kale, D.; Saeed, M.; et al.. "Do no harm: a roadmap for responsible machine learning for health care", *Nature medicine*, vol. 25–9, 2019, pp. 1337–1340.

[106] Wiggers, A.; Hoogeboom, E. “Predictive sampling with forecasting autoregressive models”. In: International Conference on Machine Learning, 2020, pp. 10260–10269.

[107] Wouters, O. J.; McKee, M.; Luyten, J. “Estimated research and development investment needed to bring a new medicine to market, 2009-2018”, *Jama*, vol. 323–9, 2020, pp. 844–853.

[108] Wu, J.; Fu, R.; Fang, H.; Zhang, Y.; Yang, Y.; Xiong, H.; Liu, H.; Xu, Y. “Medsegdiff: Medical image segmentation with diffusion probabilistic model”, *arXiv preprint arXiv:2211.00611*, 2022.

[109] Wu, K. E.; Yost, K. E.; Chang, H. Y.; Zou, J. “Babel enables cross-modality translation between multiomic profiles at single-cell resolution”, *Proceedings of the National Academy of Sciences*, vol. 118–15, 2021, pp. e2023070118.

[110] Xu, J.; Kim, S.; Song, M.; Jeong, M.; Kim, D.; Kang, J.; Rousseau, J. F.; Li, X.; Xu, W.; Torvik, V. I.; et al.. “Building a pubmed knowledge graph”, *Scientific data*, vol. 7–1, 2020, pp. 205.

[111] Yang, Z.; Zhan, F.; Liu, K.; Xu, M.; Lu, S. “Ai-generated images as data source: The dawn of synthetic era”, *arXiv preprint arXiv:2310.01830*, 2023.

[112] Yu, K.-H.; Beam, A. L.; Kohane, I. S. “Artificial intelligence in healthcare”, *Nature biomedical engineering*, vol. 2–10, 2018, pp. 719–731.

[113] Zhang, K.; Yu, J.; Yan, Z.; Liu, Y.; Adhikarla, E.; Fu, S.; Chen, X.; Chen, C.; Zhou, Y.; Li, X.; et al.. “Biomedgpt: A unified and generalist biomedical generative pre-trained transformer for vision, language, and multimodal tasks”, *arXiv preprint arXiv:2305.17100*, 2023.

[114] Zhang, R.; Isola, P.; Efros, A. A.; Shechtman, E.; Wang, O. “The unreasonable effectiveness of deep features as a perceptual metric”. In: Proceedings of the IEEE conference on computer vision and pattern recognition, 2018, pp. 586–595.

[115] Zhang, X.; Lin, D.; Pforsich, H.; Lin, V. W. “Physician workforce in the united states of america: forecasting nationwide shortages”, *Human resources for health*, vol. 18–1, 2020, pp. 1–9.

[116] Zhao, Q.; Adeli, E.; Pohl, K. M. “Training confounder-free deep learning models for medical applications”, *Nature communications*, vol. 11–1, 2020, pp. 6010.

[117] Zhou, Y.; Zhang, R.; Chen, C.; Li, C.; Tensmeyer, C.; Yu, T.; Gu, J.; Xu, J.; Sun, T. “Towards language-free training for text-to-image generation”. In: Proceedings of

the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2022, pp. 17907–17917.

Pontifícia Universidade Católica do Rio Grande do Sul
Pró-Reitoria de Pesquisa e Pós-Graduação
Av. Ipiranga, 6681 – Prédio 1 – Térreo
Porto Alegre – RS – Brasil
Fone: (51) 3320-3513
E-mail: propesq@pucrs.br
Site: www.pucrs.br