INSTITUTO DE PESQUISAS ENERGÉTICAS E NUCLEARES
Mestrado Profissional em Tecnologia das Radiações em Ciências da Saúde

PROPOSTA DE METODOLOGIA PARA AVALIAÇÃO DA APLICAÇÃO DE ALGORITMOS DE IA NO PLANEJAMENTO RADIOTERÁPICO

RENATA MENEZES LOURENÇO

Dissertação apresentada como parte dos requisitos para obtenção do Grau de Mestre Profissional em Tecnologia das Radiações em Ciência da Saúde na Área de Concentração de Processo de Radiação na Saúde

Orientadora:
Prof. Dra. Lorena Pozzo

São Paulo
2025

INSTITUTO DE PESQUISAS ENERGÉTICAS E NUCLEARES
Mestrado Profissional em Tecnologia das Radiações em Ciências da Saúde

PROPOSTA DE METODOLOGIA PARA AVALIAÇÃO DA APLICAÇÃO DE ALGORITMOS DE IA NO PLANEJAMENTO RADIOTERÁPICO

RENATA MENEZES LOURENÇO

Dissertação apresentada como parte dos requisitos para obtenção do Grau de Mestre Profissional em Tecnologia das Radiações em Ciências da Saúde na Área de Concentração de Processos de Radiação na Saúde

Orientadora:
Prof. Dra. Lorena Pozzo

São Paulo
2025

FICHA CATOLOGRÁFICA

Fonte de Financiamento: Comissão Nacional de Energias Nucleares (CNEN)

Como citar:

LOURENCO, R. M. ***PROPOSTA DE METODOLOGIA PARA AVALIAÇÃO DA APLICAÇÃO DE ALGORITMOS DE IA NO PLANEJAMENTO RADIOTERÁPICO***. 2025. 88 f. Dissertação (Mestrado Profissional em Tecnologia das Radiações em Ciências da Saúde), Instituto de Pesquisas Energéticas e Nucleares, IPEN-CNEN, São Paulo. Disponível em: <http://repositorio.ipen.br/> (data de consulta no formato: dd/mm/aaaa)

Ficha catalográfica elaborada pelo Sistema de geração automática da Biblioteca IPEN,

com os dados fornecidos pelo(a) autor(a).

Lourenco, Renata Menezes
PROPOSTA DE METODOLOGIA PARA AVALIAÇÃO DA APLICAÇÃO DE ALGORITMOS DE IA NO PLANEJAMENTO RADIOTERÁPICO / Renata Menezes Lourenco; orientadora Lorena Pozzo. -- São Paulo, 2025.
88 f.

Dissertação (Mestrado Profissional) - Programa de Pós-Graduação em Tecnologia das Radiações em Ciências da Saúde (Processos de Radiação na Saúde) -- Instituto de Pesquisas Energéticas e Nucleares, São Paulo, 2025.

FOLHA DE APROVAÇÃO

Autora: Renata Menezes Lourenço

Título: Proposta de metodologia para avaliação da aplicação de algoritmos de IA no planejamento radioterápico.

Dissertação apresentada como parte dos requisitos para obtenção do Grau de Mestre em Tecnologia das Radiações em Ciência da Saúde na Área de Avaliação de Tecnologias da Saúde - ATS.

Banca examinadora

Prof. Dr.:_____

Instituição: _____ Julgamento_____

Prof. Dr.: _____

Instituição: _____ Julgamento_____

Prof. Dr.: _____

Instituição: _____ Julgamento_____

AGRADECIMENTOS

A Deus, por orquestrar com infinita sabedoria cada detalhe, permitindo a superação de situações que, a princípio, nem eu mesma acredito que sejam transponíveis ou que eu consiga lidar. E, quando me dou conta, vejo-me seguindo os versos de Guimarães Rosa e oferecendo à vida o que ela espera de mim: coragem. Uma coragem que só floresce pela fé de que Ele estará sempre ao meu lado.

Ao meu marido, que, em sua mais genuína essência humana, oscila entre a compreensão e a exaustão. Resiliente, ele suporta e, sem reservas, expressa suas orações pelo término deste ciclo. Nos momentos de cansaço, reclama e esbraveja, mas logo se arrepende. E, sem precisar de palavras, com um olhar e o esboço de um sorriso, diz sem falar: *"Estarei sempre aqui, não importa o que aconteça."* Quero que saiba que vejo, reconheço e sou profundamente grata pela paciência, pelo suporte e pela compreensão diante da minha ausência nas rotinas do dia a dia e nos momentos de confraternização.

À minha orientadora, por quem nutro imenso respeito e admiração, uma referência marcante na minha trajetória acadêmica e pessoal. Agradeço por acreditar no meu potencial e por sua admirável inteligência emocional ao ajustar sua abordagem às particularidades do meu momento. Por me desafiar, ouvir, acolher, encorajar, apoiar, além de viabilizar conexões essenciais com pessoas e instituições fundamentais para o meu desenvolvimento. A você, *Dra. Lorena*, minha eterna gratidão.

A todos os profissionais que atuam nos bastidores cuidando de cada detalhe imprescindível para o bom andamento e o êxito de nossas atividades. A dedicação e a cordialidade, seja nos cuidados com os pequenos detalhes ou no apoio emocional durante os momentos mais desafiadores fazem a diferença na caminhada. Gostaria de mencionar alguns nomes que representam todos os que ocupam essa posição: Andrea, Sra. Salete, Jaque, Tamires e Tânia. Meu sincero agradecimento pelo carinho e pela presença de vocês.

RESUMO

O aumento da incidência global dos casos de câncer, previsto em 24 milhões de casos até 2030 destaca a urgência de ampliar o acesso à radioterapia. A disparidade é evidente: países desenvolvidos têm uma máquina para 120.000 habitantes, enquanto no Brasil, por exemplo, a proporção é de uma para 543.068. A medicina personalizada, que exige replanejamentos frequentes, tende a agravar esse cenário. A inteligência artificial surge como uma solução promissora, mas enfrenta desafios como a falta de validação e evidências robustas para sua adoção clínica.

Este trabalho apresenta uma proposta de metodologia para avaliar algoritmos de IA em radioterapia registrada no PROSPERO sob o número CRD42024574448. A avaliação engloba tanto o desempenho, considerando métricas de segmentação e planejamento dosimétrico, quanto a viabilidade clínica, analisando aspectos como dados, treinamento, rastreabilidade, reprodutibilidade e explicabilidade. O protocolo

Durante a elaboração deste documento, foram identificadas questões estruturais relevantes que antecedem a IA. Essas questões referem-se à falta de padronização na prescrição e relato de doses entre as múltiplas diretrizes clínicas bem como a existência de múltiplas diretrizes de contorno, dificultando a comparabilidade entre estudos nessas áreas. O avanço tecnológico traz novos desafios, como a ausência de conjuntos de dados padronizados para validar métricas matemáticas empregadas na avaliação da segmentação, além da crescente demanda por maior explicabilidade algorítmica.

Para garantirmos que o uso da IA possa contribuir com a redução da disparidade de acesso é necessário um esforço interdisciplinar para viabilizar a padronização de diretrizes clínicas, a criação de bancos de dados compartilhados e o desenvolvimento de algoritmos transparentes e reprodutíveis. Apesar do estágio incipiente desses sistemas na radioterapia, avanços significativos já foram alcançados. Visando a validação desses sistemas bem como a aceleração dos avanços, a padronização de métricas e ferramentas é essencial para evitar inconsistências que prejudiquem a aplicação prática.

Palavras chave: inteligência artificial, radioterapia, segmentação de tumores, planejamento dosimétrico e explicabilidade.

ABSTRACT

The increasing global incidence of cancer cases, projected to reach 24 million by 2030, highlights the urgent need to expand access to radiotherapy. The disparity is evident: developed countries have one radiotherapy machine for every 120,000 inhabitants, while in Brazil, for example, the ratio is one per 543,068. Personalized medicine, which requires frequent replanning, tends to exacerbate this scenario. Artificial intelligence (AI) emerges as a promising solution but faces challenges such as the lack of validation and robust evidence for its clinical adoption.

This study proposes a methodology to evaluate AI algorithms in radiotherapy registered in PROSPERO under the number CRD42024574448. The evaluation encompasses both performance, considering segmentation and dosimetric planning metrics, and clinical feasibility, analyzing aspects such as data, training, traceability, reproducibility, and explainability.

During the preparation of this document, relevant structural issues preceding AI were identified. These issues refer to the lack of standardization in dose prescription and reporting among multiple clinical guidelines, as well as the existence of various contouring guidelines, which hinders the comparability of studies in these areas. Technological advancements bring new challenges, such as the absence of standardized datasets for validating mathematical metrics used in segmentation evaluation, in addition to the growing demand for greater algorithmic explainability.

To ensure that AI can contribute to reducing access disparities, an interdisciplinary effort is required to enable the standardization of clinical guidelines, the creation of shared data repositories, and the development of transparent and reproducible algorithms. Despite the incipient stage of these systems in radiotherapy, significant progress has already been made. To validate these systems and accelerate advancements, the standardization of metrics and tools is essential to avoid inconsistencies that could undermine practical application.

Keywords: artificial intelligence, radiotherapy, tumor segmentation, dosimetric planning, and explainability.

LISTA DE FIGURAS

LISTA DE TABELAS

LISTA DE SIGLAS

AAPM - *American Association of Physicists in Medicine* ou Associação Americana de Físicos em Medicina

AF - Aprendizado Federado

AHRQ - *Agency for Healthcare Research and Quality* ou Agência para Pesquisa e Qualidade em Cuidados de Saúde

AIRO - *Associazione Italiana di Radioterapia e Oncologia Clinica* ou Associação Italiana de Radioterapia e Oncologia Clínica

AOCR - *Asian Oceanian Congress of Radiology* ou Congresso de Radiologia da Ásia e Oceania

ASTRO - *American Society for Radiation Oncology* ou Sociedade Americana de Oncologia por Radiação

AVD - *Average Hausdorff Distance* ou Distância Média de Hausdorff

CI - *Conformity Index* ou Índice de Conformidade

CLAIM - *Checklist for Artificial Intelligence in Medical Imaging* ou Lista de Verificação para Inteligência Artificial em Imagem Médica

CLEAR - *CheckList for EvaluAtion of Radiomics Research* ou Lista de Verificação para Avaliação de Pesquisas em Radiômica

CNN - *Convolutional Neural Network* ou Rede Neural Convolucional

CRD - *Centre for Reviews and Dissemination* ou Centro de Revisões e Disseminação

CTV - *Clinical Target Volume* ou Volume Alvo Clínico

DAHANCA - *Danish Head and Neck Cancer Group* ou Grupo Dinamarquês de Câncer de Cabeça e Pescoço

DARPA - *Defense Advanced Research Projects Agency* ou Agência de Projetos de Pesquisa Avançada de Defesa

DEGRO - *Deutsche Gesellschaft für Radioonkologie* ou Sociedade Alemã de Radioterapia e Oncologia

DFFNN – *Deep Feedforward Neural Network* ou Rede Neural de Alimentação Direta Profunda

DL - *Deep Learning* ou Aprendizado Profundo

DoD - *Department of Defense* ou Departamento de Defesa

DSC - *Dice Similarity Coefficient* ou Coeficiente de Similaridade de Dice

DVH - *Dose-Volume Histogram* ou Histograma de Dose-Volume

EMTREE - *EMBASE Medical Thesaurus* ou Tesauro Médico do EMBASE

EQUATOR - *Enhancing the Quality and Transparency of Health Research* ou Aprimorando a Qualidade e a Transparência da Pesquisa em Saúde

ESTRO - *European Society for Radiotherapy and Oncology* ou Sociedade Europeia de Radioterapia e Oncologia

FDA - *Food and Drug Administration* ou Administração de Alimentos e Medicamentos

FNN - *Feedforward Neural Network* ou Rede Neural de Alimentação Direta

FUTURE-AI - *Fairness, Universality, Traceability, Usability, Robustness and Explainability* ou Imparcialidade, Universalidade, Rastreabilidade, Usabilidade, Robustez e Explicabilidade.

GAN - *Generative Adversarial Network* ou Rede Generativa Adversária

GHG - *Global Quality Assurance of Radiation Therapy Clinical Trials Harmonisation Group* ou Grupo de Harmonização Global de Garantia de Qualidade em Ensaios Clínicos de Radioterapia

GI - *Gradient Index* ou índice de gradiente

GMLP - *Good Machine Learning Practice for Medical Device Development: Guiding Principles* ou Boas Práticas de Aprendizado de Máquina para o Desenvolvimento de Dispositivos Médicos: Princípios Orientadores

GNN - *Graph Neural Network* ou Rede Neural de Grafos

GORTEC - *Groupe Oncologie Radiothérapie Tête et Cou* ou Grupo de Oncologia e Radioterapia de Cabeça e Pescoço

GRADE - *Grading of Recommendations Assessment, Development, and Evaluation* ou Graduação das Recomendações de Avaliação e Desenvolvimento

Grad-CAM - *Gradient-weighted Class Activation Mapping* ou Mapeamento de Ativação de Classe Ponderado por Gradiente

HD - *Hausdorff Distance* ou Distância de Hausdorff

HI - *Homogeneity Index* ou índice de homogeneidade

HTA - *Health Technology Assessment* ou Avaliação de Tecnologias em Saúde

IA - Inteligência Artificial

IARC - *International Agency for Research on Cancer* ou Agência Internacional de Pesquisa sobre o Câncer

ICRU - *International Commission on Radiation Units and Measurements* ou Comissão Internacional de Unidades e Medidas de Radiação

IMRT - *Intensity-Modulated Radiation Therapy* ou Terapia de Radiação com Intensidade Modulada

ISOMAP – *Isometric Mapping* ou Mapeamento Isométrico

KPCA – *Kernel Principal Component Analysis* ou Análise de Componentes Principais com Kernel

LIME - *Local Interpretable Model-agnostic Explanation* ou Explicação Local Interpretável Independente de Modelo

LLE - *Locally Linear Embedding* ou Incorporação Linear Local

MC - Monte Carlo

MDS – *Multidimensional Scaling* ou Escalonamento Multidimensional

MeSH - *Medical Subject Headings* ou Cabeçalho de Assuntos Médicos

MIS - *Medical Image Segmentation* ou Segmentação de Imagens Médicas

MINIMAR - *MINimum Information for Medical AI Reporting* ou Informação Mínima para Relatórios de IA Médica

MI-CLAIM - *Minimum information about Clinical Artificial Intelligence Modeling* ou Informação Mínima sobre Modelagem de Inteligência Artificial Clínica

ML - *Machine Learning* ou Aprendizado de Máquina

MLops - *Machine Learning Operations* ou Operações de Aprendizado de Máquina

MHRA - *Medicines and Healthcare Products Regulatory Agency* ou Agência Reguladora de Medicamentos e Produtos de Saúde

OAR - *Organs at Risk* ou Órgãos de Risco

OMS - Organização Mundial da Saúde

PET-CT - *Positron Emission Tomography - Computed Tomography* ou Tomografia por Emissão de Pósitrons – Tomografia Computadorizada

PRISMA - *Preferred Reporting Items for Systematic Reviews and Meta-Analyses* ou Itens Preferidos para Relato de Revisões Sistemáticas e Meta-Análises

PRISMA-AI - *Preferred Reporting Items for Systematic Reviews and Meta-Analyses tailored for Artificial Intelligence* ou Itens Preferenciais de Relato para Revisões Sistemáticas e Meta-análises adaptados para Inteligência Artificial

PRISMA-P - *Preferred Reporting Items for Systematic Reviews and Meta-Analyses Protocols* ou Itens de Relato Preferenciais para Protocolos de Revisões Sistemáticas e Meta-análises.

PROBAST - *Prediction Model Risk of Bias Assessment Tool* ou Ferramenta de Avaliação para Risco de Viés em Modelos de Predição

PROBAST+AI - *Prediction model Risk Of Bias Assessment Tool for Artificial Intelligence* ou Ferramenta de Avaliação para Risco de Viés em Modelos de Predição em Inteligência Artificial

ProtoPNet - *Prototypical Part Network* ou Rede de Partes Prototípicas

PROSPERO - *International Prospective Register of Systematic Reviews* ou Registro Internacional Prospectivo de Revisões Sistemáticas

PRV - *Planning Risk Volume* ou Volume de Risco Planejado

PTV - *Planned Target Volume* ou Volume Alvo Planejado

RATING - *Radiotherapy Treatment plannINg study Guidelines* ou Estudos de Planejamento de Tratamento em Radioterapia

RF - *Random Forest*

RM - Ressonância Magnética

RNA – Rede Neural Artificial

RNN - *Recurrent Neural Network* ou Rede Neural Recorrente

RPA - *Radiotherapy Planning Assistant* ou Assistente de Planejamento de Radioterapia

RTOG - *Radiation Therapy Oncology Group* ou Grupo de Oncologia de Terapia por Radiação

RTTQA - *Radiotherapy Trials Quality Assurance* ou Garantia de Qualidade dos Ensaios de Radioterapia

SHAP - *Shapely Additive Explanations* ou Explicações Aditivas de Shapley

SBRT - Sociedade Brasileira de Radioterapia

SBRT - *Stereotactic Body Radiotherapy* ou Radiocirurgia Corporal Estereotática

SRS - *Stereotactic Radiosurgery* ou Radiocirurgia Estereotática

SRT - *Stereotactic Radiotherapy* ou Radioterapia Estereotática

SVM - *Support Vector Machine* ou Máquina de Vetores de Suporte

TC - Tomografia Computadorizada

TCAV - *Testing with Concept Activations Vectors* ou Testando com Vetores de Ativação de Conceitos

TRIPOD - *Transparent Reporting of a Multivariable Prediction Model for Individual Prognosis or Diagnosis* ou Relato Transparente de um Modelo de Predição Multivariável para Prognóstico ou Diagnóstico Individual.

t-SNE - *t-Distributed Stochastic Neighbor Embedding* ou Incorporação Estocástica de Vizinhos Distribuída em t

XAI - *Explainable Artificial Intelligence* ou Inteligência Artificial Explicável

SUMÁRIO

1. INTRODUÇÃO

A incidência global de câncer está aumentando, com a **IARC** (*International Agency for Research on Cancer* ou Agência Internacional de Pesquisa sobre o Câncer) projetando 24 milhões de novos casos até 2030. No Brasil, são esperados cerca de 640 mil casos de câncer até 2030, com mais da metade necessitando de radioterapia. O Brasil já enfrenta desafios significativos na infraestrutura de radioterapia. De acordo com dados do relatório da **SBRT** (Sociedade Brasileira de Radioterapia), em 2020, havia uma máquina disponível para cada 543.068 habitantes (1), em comparação com, aproximadamente, uma para cada 120.000 em países de alta renda (2).

Além do aumento da incidência de câncer, outro fator que pode impactar a capacidade de atendimento das unidades de radioterapia é a medicina personalizada, que exigirá cada vez mais planos de tratamento adaptáveis, que reavaliem sistematicamente a localização, forma e volume do tumor, ajustando a distribuição de dose para incorporar variações anatômicas e de posicionamento no plano de tratamento. Isso implica maior complexidade e mais tempo para realizar o replanejamento ao longo do tratamento.

Nas últimas décadas, houve muitos avanços na radioterapia, como orientação por imagem tridimensional, modulação de intensidade e robótica, que contribuíram tanto para maior precisão do tratamento quanto para a redução da toxicidade nos tecidos saudáveis. Seguindo essa trajetória evolutiva, com o avanço da inteligência artificial (IA), cresce a expectativa de que ela tenha o potencial de remodelar o campo da radioterapia, colaborando com médicos e físicos médicos para aumentar a produtividade através da otimização dos planos de tratamento (3).

Ainda no campo da terapia por radiação, melhorar a produtividade permitiria ganhos que vão além do aspecto operacional, impactando o desfecho clínico. Visto que a eficiência operacional permite o tratamento de mais pacientes e já está comprovado que adiar o tratamento tem impacto negativo no controle e no prognóstico devido à progressão da doença (4).

Portanto, ferramentas que possam contribuir para aumentar a produtividade sem comprometer a *segurança* e a eficácia do tratamento são alternativas para expandir o acesso ao tratamento. Vale destacar que, segundo a **OMS** (Organização Mundial da Saúde), o termo segurança do paciente significa "a redução do risco de dano desnecessário associado ao cuidado de saúde para um mínimo aceitável". Um mínimo aceitável refere-se às noções coletivas dadas ao conhecimento atual, recursos disponíveis e o contexto em que o cuidado foi oferecido, ponderados contra o risco de não tratamento ou outro tratamento" (5).

Nesse contexto, a consolidação de evidências é fundamental para validar e assegurar a performance, robustez e transparência necessárias para viabilizar a adoção de algoritmos de IA em aplicações clínicas críticas conciliando as demandas por personalização do tratamento, eficiência operacional e uso racional dos recursos disponíveis.

Em sintonia com essa necessidade, impulsionados pelos avanços dos sistemas de IA, especialmente na segmentação de tumores, observa-se um crescente aumento na publicação de estudos. Concomitantemente, cresce, também, o número de publicações orientadas a garantir a robustez e transparência dos relatos e dos algoritmos de IA em saúde. Essas publicações vão desde ferramentas para avaliar criticamente as aplicações dos algoritmos passando por diretrizes/princípios orientadores para direcionar o desenvolvimento de software (6) até diretrizes conjuntas de sociedades científicas orientadas a validação clínica (7).

Essa necessidade de municiar a comunidade científica com recursos para avaliações mais rigorosas das aplicações de IA em saúde já havia sido suscitada numa revisão seminal conduzida pela revista científica "The Lancet Digital Health" em 2019 a qual revelou que menos de 0,1% dos estudos sobre IA em imagens médicas atendiam aos padrões de qualidade para adoção clínica (8).

Atualmente, na área da saúde existem mais de 20 ferramentas para avaliação de IA (9). A expectativa é que esse número continue a aumentar. Um exemplo, é o **PROBAST+AI** (*Prediction model Risk Of Bias Assessment Tool for Artificial Intelligence* ou Ferramenta de Avaliação para Risco de Viés em Modelos de Predição em Inteligência Artificial) que está em processo de desenvolvimento (10). Essa

variedade de ferramentas demanda uma análise criteriosa por parte dos pesquisadores. A seleção da diretriz mais adequada deve levar em conta o tema central da pesquisa e a compatibilidade entre os conteúdos. Em alguns casos, a combinação de diferentes diretrizes pode ser necessária para abarcar a complexidade do tema (6).

A complexidade vai além da escolha da ferramenta. A diversidade de métricas empregadas em estudos de segmentação tumoral (11) e cálculo de distribuição de dose (12) adiciona um nível extra de dificuldade às revisões sistemáticas, tornando a comparação entre os resultados mais desafiadora e, consequentemente, comprometendo a robustez das evidências científicas. Essa heterogeneidade metodológica representa um desafio significativo para avaliação da IA como uma possível solução para expansão do acesso aos tratamentos.

Tendo em vista esse panorama, o intuito deste trabalho é apresentar uma proposta de metodologia para a avaliação da aplicação de algoritmos de IA no planejamento radioterápico que abarque tanto a perspectiva do desempenho quanto da aplicação na prática clínica. Para isso, no âmbito do desempenho, serão analisadas as múltiplas métricas utilizadas na segmentação e cálculo de dose com o intuito de embasar uma recomendação que viabilize a comparabilidade entre os estudos. Já na perspectiva de implementação clínica, serão avaliados critérios que permitam dar visibilidade sobre questões como:

> *Dados*: a diversidade e a qualidade dos dados são essenciais para garantir a generalização do modelo (13). Ou seja, um mix de dados de imagem representativo que englobe múltiplos centros de saúde, diferentes fornecedores e contextos com recursos limitados, como os encontrados em países de baixa e média renda (14). Esse último elemento é crucial para garantia da equidade de acesso à tecnologia.
>
> *Treinamento*: área crítica para assegurar a confiabilidade da IA, pois impacta diretamente os resultados do modelo e é uma das principais causas de viés (14).

Rastreabilidade: a documentação completa do processo de desenvolvimento e funcionamento permite identificar as razões de decisões errôneas, evitando erros futuros (14).

Robustez: entendida como a capacidade da IA em manter a exatidão dos resultados em condições variáveis do mundo real. No caso das imagens médicas, a heterogeneidade pode afetar o desempenho do algoritmo (14).

Reprodutibilidade: A disponibilização dos meios que permitam reproduzir de modo exato o processo de geração dos resultados possibilita a realização de testes em outras coortes (15).

Explicabilidade: resultados compreensíveis para uma audiência específica.(16).

1.1. OBJETIVO

Propor uma metodologia para avaliar a aplicação de algoritmos de IA no planejamento radioterápico. A avaliação abrangerá a segmentação de volumes-alvos e órgãos de risco, bem como o planejamento dosimétrico, mapa de distribuição de dose. Para elaboração do documento será utilizada a diretriz de recomendação para a elaboração de protocolos de revisões sistemáticas e meta-análises, **PRISMA-P** (*Preferred Reporting Items for Systematic Reviews and Meta-Analyses Protocols* ou Itens de Relato Preferenciais para Protocolos de Revisões Sistemáticas e Meta-análises). A análise se concentrará no desempenho e na viabilidade de adoção destes algoritmos na prática assistencial em tumores de cabeça e pescoço, mama e próstata. A definição dos sítios tumorais foi baseada nos critérios de complexidade e incidência.

1.2. JUSTIFICATIVA

A apresentação de um protocolo pode contribuir com a estruturação de estudos que forneçam dados robustos e confiáveis a fim de viabilizar uma análise objetiva e criteriosa dos desfechos obtidos. Evidências bem documentadas fortalecem a credibilidade das pesquisas e facilitam a aceitação dos resultados pela comunidade científica. Além disso, ao fornecer dados consistentes e transparentes, os estudos

permitem que outros pesquisadores reproduzam as descobertas e verifiquem a validade das conclusões.

Sob a ótica dos desenvolvedores, o intuito é promover a pesquisa e o relato de mecanismos que viabilizem a generalização, robustez, rastreabilidade, reprodutibilidade e compreensão dos algoritmos com a finalidade de mitigar as barreiras de adoção.

Tendo em vista a infraestrutura limitada e as complexidades para expansão dos serviços de radioterapia nos países de baixa e média renda, é fundamental que as pesquisas em IA, também, sejam direcionadas para o desenvolvimento de tecnologias que promovam um acesso mais equânime reduzindo as desigualdades entre as diversas regiões.

2. FUNDAMENTOS CONCEITUAIS

Antes de adentrar na fase de desenvolvimento deste trabalho é importante estabelecer uma base conceitual sobre inteligência artificial, sua aplicação na radioterapia e os desafios para adoção na prática clínica.

2.1. DA INTELIGÊNCIA ARTIFICIAL AO APRENDIZADO PROFUNDO DE MÁQUINA

A inteligência artificial é o campo da ciência da computação que visa permitir que computadores ou robôs controlados por computadores possam desenvolver tarefas que remetam ao comportamento humano, como por exemplo, aprender com experiência passada, reconhecer padrões e tomar decisões. Na **Figura 1** (17), uma representação gráfica das múltiplas áreas da IA.

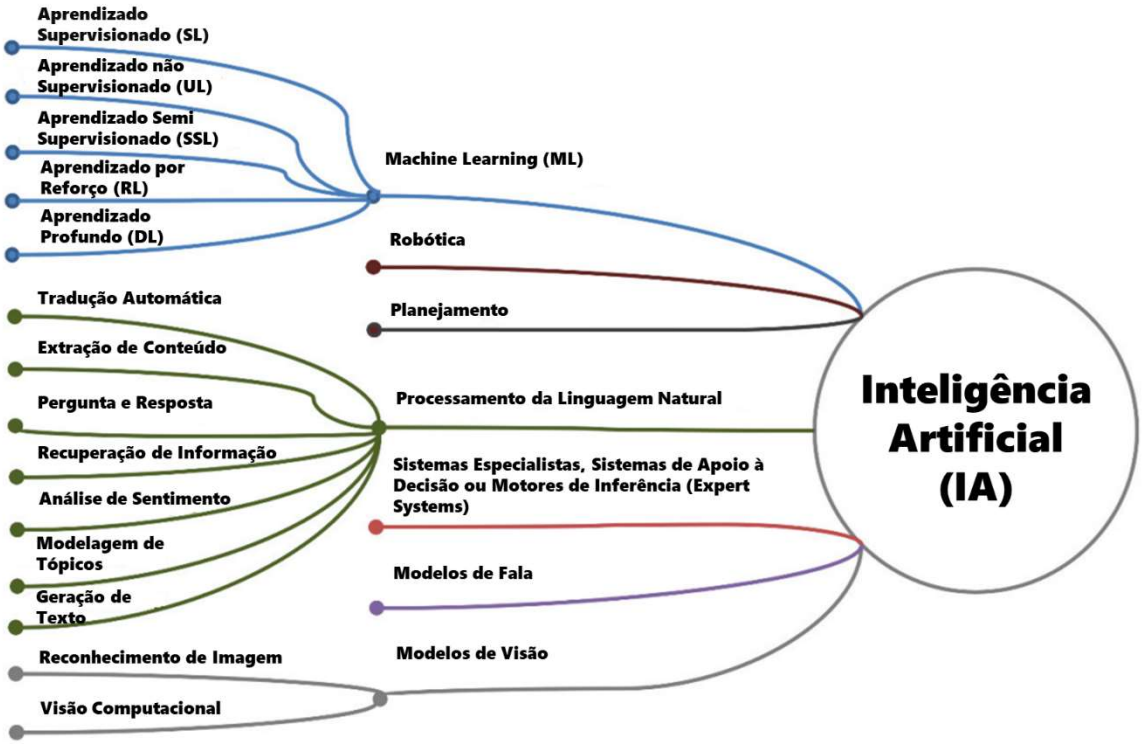


Figura 1: Subseções de inteligência artificial (17)

O aprendizado de máquina, doravante denominado *Machine Learning* ou simplesmente ML, é um subcampo da IA que materializa o potencial inerente a ideia de IA. Ou seja, propicia o desenvolvimento de sistemas capazes de aprender com os dados permitindo que computadores adquiram conhecimento a partir de experiências retrospectivas, adaptando-se e aprimorando habilidades para realização de diversas tarefas (17).

Em linhas gerais, os modelos de ML podem ser classificados em clássicos e modernos conforme ilustrados na **Figura 2** (18).

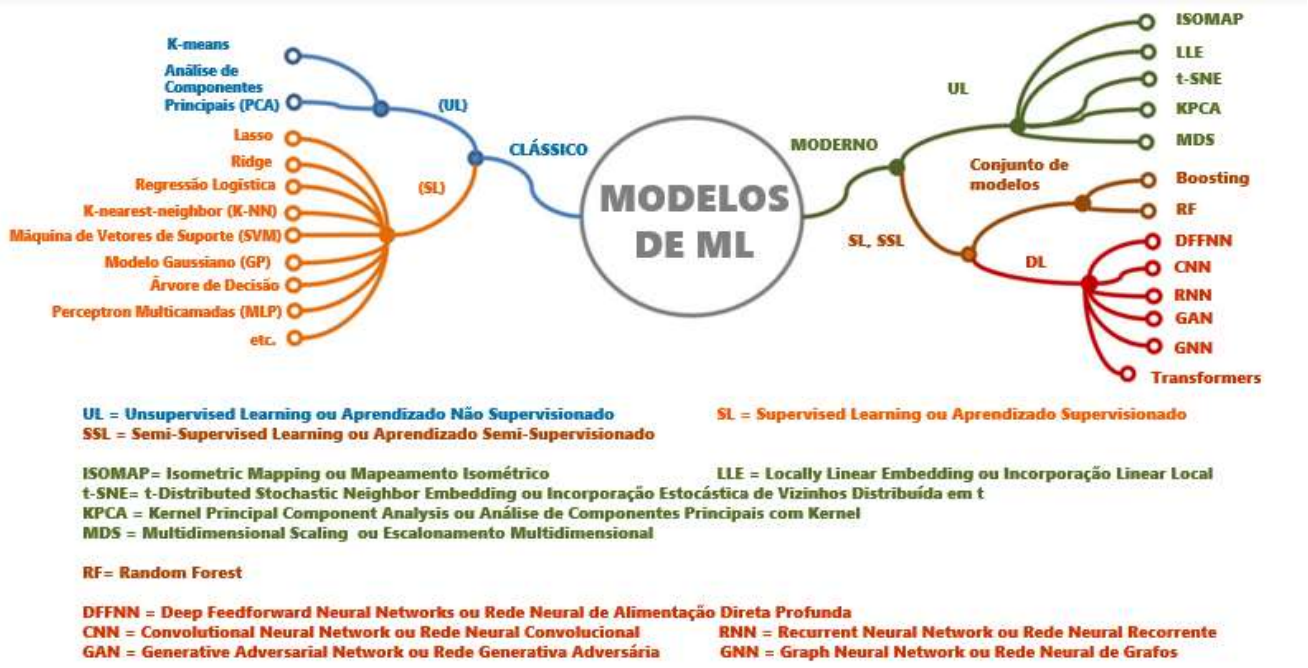


Figura 2: Modelos de aprendizado de máquina clássicos e modernos (18)

Uma vez que os dados estejam preparados, a escolha do algoritmo dependerá de vários fatores, incluindo:

1. tamanho, qualidade e natureza dos dados do domínio;
2. tempo computacional disponível;
3. urgência da tarefa; e
4. estrutura das previsões desejadas e a função de perda a ser minimizada (19).

A aprendizagem profunda, doravante denominada *Deep Learning* ou simplesmente DL, é um subcampo do ML, mais especificamente, um subcampo de um modelo conhecido como rede neural artificial (RNA). As RNAs são modelos computacionais inspirados no funcionamento do cérebro humano que possuem unidades (neurônios) organizadas em camadas. Essas camadas são organizadas em três seções, são elas: a camada de entrada, que recebe os dados; a(s) camada(s) oculta(s), que processa(m) os dados e extrai(em) as características relevantes; e a camada de saída, que produz a resposta final do modelo (19).

As RNAs podem ser divididas em redes rasas (“*shallow neural networks”* com uma camada oculta) e DL (com mais camadas ocultas). O número de camadas está diretamente relacionado à capacidade preditiva das redes neurais (19). Além do

número de camadas, as principais diferenças entre DL e RNA estão relacionadas às conexões (modo como os neurônios artificiais são organizados e interligados) e a capacidade de abstração (aprender características mais complexas e abstratas). De fato, as RNAs tradicionais geralmente são limitadas a três camadas e, em geral, têm dificuldade em manter a precisão evitando erros significativos em dados desconhecidos ou fora da amostra de treinamento. Visto que, os dados brutos são transformados e organizados internamente para atender uma tarefa específica, não capturando abstrações amplas ou universais (20).

A principal vantagem do DL está relacionada à capacidade de resolver tarefas utilizando o método "*end-to-end*". Esse método refere-se à habilidade de automatizar todas as etapas do processo de aprendizado, desde a entrada até a saída, sem precisar de intervenções humanas para pré-definir as características ou etapas intermediárias. Essa abordagem reduz as exigências em relação a preparação dos dados em formatos específicos para que o modelo possa processá-los. Uma vez que, o modelo aprende automaticamente as características importantes e relaciona essas características com a variável alvo identificando de modo independente padrões ou relações presentes nos dados é possível prever ou classificar a saída com base nas entradas (18).

A rede realiza o processo complexo e trabalhoso de seleção das características significativas, o que simplifica consideravelmente o trabalho do desenvolvedor. No entanto, é importante ressaltar que a qualidade e a representatividade dos dados de treinamento são fatores cruciais para o desempenho do modelo. Quanto mais diversos e relevantes forem os dados, maior será a capacidade do modelo de generalizar para novos exemplos (18).

A vasta gama de arquiteturas de DL pode ser agrupada em três categorias básicas, conforme relatado a seguir (18):

1. Rede neural padrão de alimentação direta (FFNN);
2. Rede neural recorrente (RNN);
3. Rede neural convolucional (CNN);
4. Arquiteturas híbridas, que incluem elementos das três arquiteturas básicas 1, 2 e 3 como redes siamesas e “*transformers*”.

Apesar da RNA e do DL serem originários das décadas de 1940 e 1960, respectivamente, foi apenas em 1990 que surgiu a primeira aplicação prática com o desenvolvimento do LeNet (uma das primeiras redes neurais convolucionais utilizada em sistemas bancários para ler automaticamente valores em cheques manuscritos). O exponencial crescimento do DL a partir de 2010 pode ser atribuído à combinação de três fatores principais (19):

1. novos avanços algorítmicos que melhoraram significativamente a precisão das aplicações e ampliaram os domínios de aplicação;
2. disponibilidade de uma enorme quantidade de dados para treinar as RNAs;
3. aumento do poder computacional.

2.2. O POTENCIAL DA IA NA OTIMIZAÇÃO DO FLUXO DE TRABALHO NA RADIOTERAPIA

O uso de IA na radioterapia está em rápido desenvolvimento e merece especial atenção devido ao potencial de impactar profundamente o tratamento dos pacientes oncológicos. Nos últimos anos, foram testemunhados avanços substanciais, particularmente na automação de contornos. A **Figura 3** (21) ilustra de modo resumido as oportunidades da aplicação de IA em cada etapa do fluxo operacional dessa área.

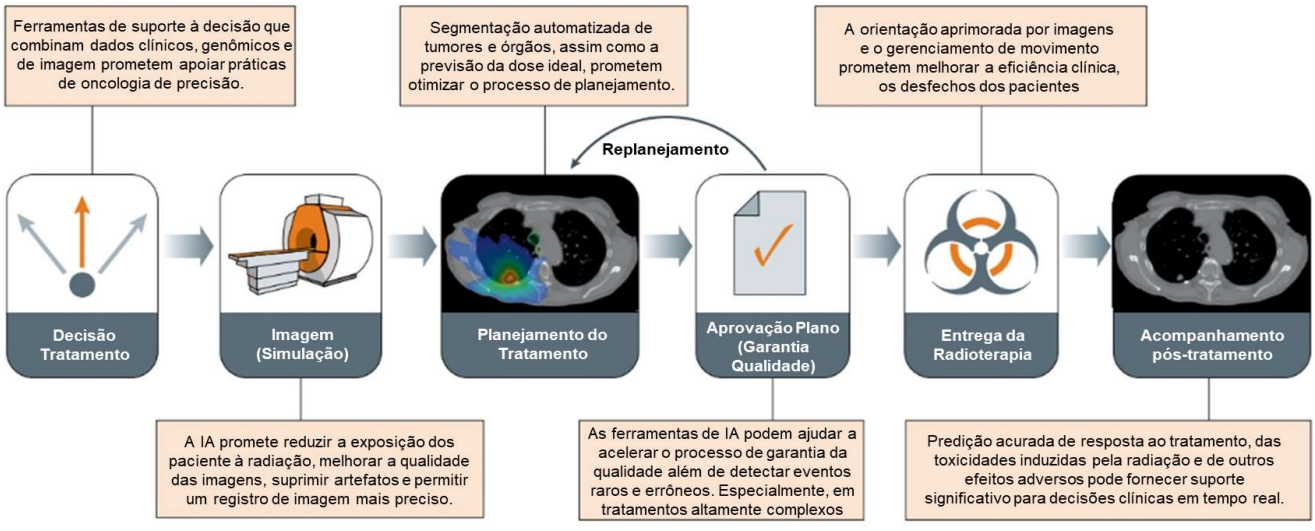


Figura 3: Aplicações de IA no fluxo da radioterapia (21)

O escopo desse trabalho se limita ao planejamento do tratamento abordando tanto a segmentação quanto o mapa de distribuição de dose. Não serão abordadas questões relativas à aplicação da IA na orientação do feixe e na variação da intensidade do feixe de radiação, mapa de fluência (22).

Quando se fala de planejamento dosimétrico, o método de Monte Carlo (MC) é considerado o padrão-ouro para cálculo de dose, pois simula, de forma precisa, o caminho das partículas ao interagirem com diferentes tipos de tecidos e materiais, considerando aspectos como absorção e espalhamento (23). A dose resultante deve ser comparada a uma dose de referência e, em caso de discrepâncias, ajustes nos parâmetros de entrada utilizados como base para a simulação devem ser realizados.

Embora seja um recurso poderoso para simulação do transporte de radiação, o método de Monte Carlo apresenta uma limitação em termos de tempo de processamento computacional. Isso acontece devido à sua natureza, que envolve várias simulações de interação da radiação com os tecidos obedecendo a funções de probabilidade conhecidas. Apesar do contínuo e expressivo progresso dos “hardwares” e da relativa facilidade do uso de técnicas como paralelismo de código, simulações complexas continuam sendo um desafio (24).

Diante desse cenário, a integração de DL com o MC surge como uma alternativa promissora para atenuar essa limitação. Essa abordagem automatiza o processo iterativo de ajuste dos parâmetros de entrada da simulação de Monte Carlo, assegurando a correspondência entre a dose calculada e a dose de referência. Graças a referida integração é possível reduzir significativamente o tempo necessário para os cálculos de dose, levando-o de horas ou dias para apenas minutos (24). Visto que, a precisão no ajuste dos parâmetros de entrada elimina a necessidade de executar inúmeras simulações até alcançar o resultado desejado. A **Figura 4** representa este processo de maneira gráfica.

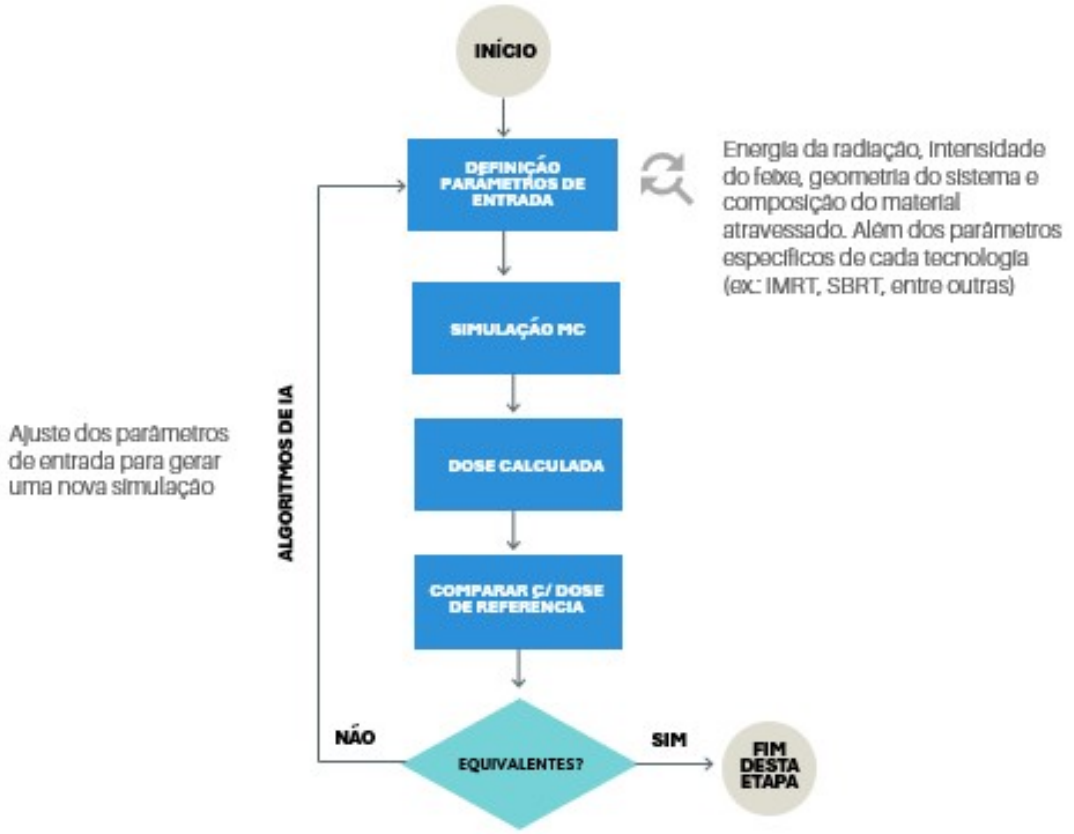


Figura 4: Cálculo de dose associando MC e DL. Fonte: Gerado pelo autor

Ainda que seja uma alternativa propícia, é importante estar atento aos desafios. As redes neurais são efetivamente agnósticas em relação à física ou ao modelo. Isto é, não possuem pressuposições específicas sobre os dados que analisam permitindo sua aplicação numa ampla variedade de tarefas, independentemente, do domínio. Simplesmente, aprendem propriedades a partir de um conjunto de dados de treinamento. Portanto, considerando que, geralmente há requisitos quantitativos associados às tarefas de simulação de MC, como a estimativa exata de energia ou dose depositada e a estimativa exata de propriedades das partículas, entre outras, os algoritmos de DL em simulações de MC precisam ser consistentemente exatos e essa exatidão precisa ser avaliada (24).

2.3. OS DESAFIOS PARA INCORPORAÇÃO NA PRÁTICA CLÍNICA

Tendo em vista os potenciais benefícios, a integração da IA na radioterapia, aparentemente, é uma tendência irreversível (25). A segmentação de tumores é, sem dúvida, a aplicação mais avançada dessa tecnologia na área (7) e um dos motivos é a sua intrínseca conexão com a radiologia, reconhecidamente a especialidade médica com mais aplicações de IA na medicina (25).

Embora em contextos distintos, tanto a radiologia quanto a radioterapia compartilham um cenário de constante evolução tecnológica, com muitos usos ainda teóricos, em desenvolvimento ou limitados a instituições individuais (13). A mudança desse paradigma possui desafios multifacetados que incluem aspectos éticos, regulatórios, jurídicos, financeiros, tecnológicos entre outros (18) (26). No âmbito desta pesquisa, dentre os vários desafios tecnológicos que permeiam esta área, serão priorizados três que são (13):

1. *Volume e qualidade dos dados* -> garantir a disponibilidade de dados que englobem múltiplas regiões geográficas, instituições e populações diversas de pacientes é fundamental. Sem isso, a cada nova instituição, é necessária uma validação externa e treinamento adicional utilizando dados com população de pacientes diversificadas (13). Considerando o cenário de desenvolvimento de algoritmos com dados unicêntricos, a *reprodutibilidade*, isto é, a disponibilização de meios que assegurem a reprodução exata do processo de geração dos resultados, se torna, ainda, mais relevante para realização de testes com outras coortes (15).

   A variabilidade e complexidade das doenças tornam a tarefa de modelagem computacional muito mais desafiadora quando comparada a outras aplicações, como reconhecimento de imagem ou fala. Sendo assim, para garantir um modelo eficaz e robusto, é necessário um volume muito maior de dados em relação a outros tipos de casos de uso (20).

   Além disso, no que diz respeito à qualidade, os dados médicos se caracterizam por serem altamente heterogêneos, ambíguos, ruidosos e incompletos, o que acrescenta mais uma peculiaridade frente a outros domínios, onde os dados

costumam ser mais limpos e bem estruturados (20). Focando especificamente em imagens médicas, essa heterogeneidade pode ser atribuída a múltiplos fatores. Com a finalidade de elucidar um pouco mais o tema, seguem dois exemplos: a) variações nos equipamentos, devido a diferenças específicas de fornecedores; b) parâmetros de aquisição de imagens que variam entre centros (resolução, espessura dos cortes e tempos de escaneamento). (14)

Tais diferenças, que refletem práticas clinicas reais, podem resultar em algoritmos que apresentam bons resultados em testes padronizados, mas desempenho insatisfatório em aplicações práticas (26). Por esse motivo, a importância de medidas preventivas e corretivas para garantir a robustez dos algoritmos frente a essas variações. Entende-se como ***robustez*** a capacidade da IA em manter a exatidão dos resultados em condições variáveis do mundo real. (14)

2. *Sistema de manutenção e supervisão auditáveis* -> com a finalidade de monitorar o desempenho da inteligência artificial ao longo do tempo evitando a deriva de conceito que nada mais é que a perda do poder preditivo ao longo do tempo devido a uma mudança gradual, cíclica ou súbita em relação a base de dados utilizada para treinamento da ferramenta(13).
   A seguir, alguns conceitos chaves para facilitar o entendimento dos diferentes tipos de deriva de conceito (27):

   - Probabilidade combinada ($P(x, y)$): descreve a probabilidade de tanto x (entrada) quanto y (saída) ocorrerem simultaneamente.
   - Probabilidade condicional $\left(P((y|x)\right)$: isso mostra a probabilidade de y ocorrer, dado que x é conhecido.
   - Ao longo do tempo (t_0 e t_1):
     - $P_{to}(x, y)$: probabilidade combinada no tempo t_0
     - $P_{t1}(x, y)$: probabilidade combinada no tempo t_1

   Se as probabilidades em t_0 e t_1 são diferentes, ocorreu uma deriva de conceito.

Existem três tipos de deriva de conceito, a saber (27):

- *Deriva de conceito virtual*: quando a probabilidade de x muda, mas a probabilidade de y dado x permanece a mesma. Ou seja: $(P_{to}(x) \neq P_{t1}(x))$ e $P_{t0}(y|x) = P_{t1}(y|x)$

  Esse tipo de deriva está associado a atualizações técnicas ou de hardware (por exemplo, novos equipamentos de imagem), que exigem calibração, mas não retreinamento do mecanismo de decisão.

  Um exemplo relacionado à segmentação de tumores seria a introdução de uma nova máquina de tomografia computadorizada (TC) produzindo imagens com diferenças sutis na resolução espacial, níveis de brilho ou contraste. Essas alterações nas características gerais das imagens que o modelo recebe como entrada, podem levar a desvios no desempenho do sistema. Nessa situação, seria necessária recalibração do modelo para corrigir as discrepâncias. A recalibração não necessariamente demanda um retreinamento do algoritmo, mas pode exigir ajustes pontuais (“fine-tuning”).

- *Deriva de conceito real*: quando a probabilidade de y dado x muda e a probabilidade de x permanece a mesma. Ou seja $P_{t0}(y|x) \neq P_{t1}(y|x)$ e $P_{to}(x) = P_{t1}(x)$

  Essa forma de deriva reflete mudanças no entendimento clínico, diretrizes ou práticas de diagnóstico, necessitando de retreinamento dos modelos para se alinhar com os novos padrões.

  Um exemplo prático em modelos de planejamento de radioterapia ocorre quando as diretrizes clínicas evoluem. Por exemplo, se novas pesquisas indicarem que a proteção de órgãos adjacentes é prioritária, mesmo para tumores similares, o modelo precisará ser retreinado. Isso porque a definição de 'tratamento ideal' muda, tornando obsoleto o critério de decisão anterior.

- *Deriva de conceito híbrida*: num ambiente aberto, a deriva conceitual real e virtual ocorrendo de modo simultâneo no fluxo de dados. Ou seja, $P_{t0}(y|x) \neq P_{t1}(y|x)$ e $P_{to}(x) \neq P_{t1}(x)$. Este é o tipo de deriva mais complexo e comum, exigindo monitoramento e adaptação robustos.

  A deriva conceitual híbrida combina mudanças de hardware e clínicas, exigindo atualizações abrangentes do modelo para lidar com mudanças tanto nos dados de entrada quanto nas interpretações médicas.

  Em um cenário prático, considere um modelo de previsão de dose utilizado em uma clínica voltada para pacientes adultos. Com a introdução de uma nova tecnologia de tratamento, como a terapia com prótons, e a expansão do atendimento para uma nova população, como pacientes pediátricos, ocorrem mudanças não apenas nas características anatômicas dos pacientes, mas também na relação entre a anatomia e a dose prescrita. Dessa forma, o modelo precisa ser ajustado para se adaptar aos novos dados e manter a precisão das previsões.

Para lidar com o desafio de manter o poder preditivo quando os novos dados não se encaixam nos padrões estabelecidos pelos dados históricos é importante monitorar, adaptar e compreender as mudanças que ocorrem nos dados e no comportamento do modelo ao longo do tempo. Para isso, cada sistema de IA deve integrar ferramentas de rastreabilidade desde o desenvolvimento permitindo monitorar o funcionamento em tempo real, registrando métricas, erros, desvios e degradação no desempenho durante sua operação. Sendo a *rastreabilidade* entendida como a documentação completa do processo de desenvolvimento e funcionamento para identificar as razões de decisões errôneas, evitando erros futuros (14). As principais estatísticas do sistema e, se possível, o feedback dos profissionais de saúde, devem ser registrados em um repositório de modelos. A ferramenta de auditoria pode ser incluída na estrutura de operações do modelo ("*framework* ***MLops*** *- Machine Learning Operations*"). Em cenários de aprendizado dinâmico, em que o

algoritmo é capaz de se adaptar aos dados de produção ou as devolutivas fornecidos por profissionais de saúde, o sistema de monitoramento deve registrar não apenas as métricas de desempenho, mas também os processos e ferramentas utilizados para o aprendizado contínuo do sistema (14). Na **Figura 5**, é apresentada uma estrutura geral para a detecção de deriva de conceito (28).

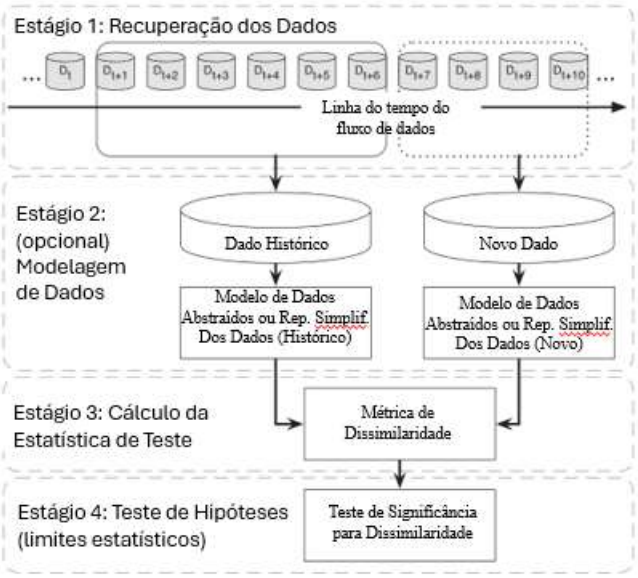


Figura 5: Uma estrutura geral para detecção de deriva de conceito (28)

3. *O efeito caixa-preta* -> refere-se a falta de transparência dos algoritmos de IA em relação à informação disponibilizada assim como do processo para atingi-la comprometendo a confiabilidade e em decorrência direta a ampla adoção em áreas de alto risco como saúde, finanças e segurança. Conforme a aplicação desses algoritmos se expande para essas áreas de alta criticidade aumenta a necessidade de se projetar sistemas que garantam a interpretabilidade e explicabilidade.

   Em 2017, ainda, no intuito de desmistificar a "caixa preta" a **DARPA** (*Defense Advanced Research Projects Agency* ou Agência de Projetos de Pesquisa Avançada de Defesa) supervisionada e financiada pelo **DoD** (*Department of Defense* ou Departamento de Defesa) nos Estados Unidos, cria o programa de **XAI** (*Explainable Artificial Intelligence* ou Inteligência Artificial Explicável). O objetivo dessa iniciativa concluída em 2021 foi de criar um conjunto de técnicas

de aprendizado de máquina novas ou modificadas que produzissem modelos explicáveis os quais combinados com técnicas de explicação eficazes viabilizassem a compreensão e confiança dos usuários finais (29). Como o próprio nome já sugere, está intimamente ligado com os estudos na área de explicabilidade.

Não existe um consenso na comunidade científica de IA em relação aos conceitos de interpretabilidade e explicabilidade o que resulta numa frequente intercambialidade entre estes termos. Para fins deste trabalho, será adotada a seguinte distinção (16):

- **Interpretabilidade:** Um modelo interpretável possui uma estrutura e um processo de tomada de decisão transparentes facilitando a compreensão de como funciona. É uma característica que está diretamente associada ao design do modelo devido à sua simplicidade ou estrutura. Para esse tipo de modelo de IA, o observador humano é capaz de compreender seu funcionamento sem necessidade de maiores explicações. Ex.: métodos clássicos de ML, como árvore de decisão (18).

- **Explicabilidade:** Diz respeito aos *métodos* utilizados para tornar um modelo de IA mais compreensível. Está frequentemente associada a técnicas aplicadas após a construção do modelo para tornar suas previsões entendíveis para uma audiência específica quando um modelo não é naturalmente interpretável como ocorre com os métodos de **SVM** (*Support Vector Machine* ou Máquina de Vetores de Suporte) e DL.

O **NIST** (*National Institute of Standards and Technology* ou Instituto Nacional de Padrões e Tecnologia), órgão norte-americano responsável por promover a inovação e a competitividade industrial, em colaboração com especialistas em IA, definiram uma base de quatro princípios para guiar o desenvolvimento de sistemas de IA explicáveis. Esses princípios, que serão detalhados a seguir, visam assegurar que as explicações geradas sejam não apenas

compreensíveis, mas também reflitam com precisão o funcionamento interno do sistema e reconheçam as limitações inerentes ao modelo (30).

- ***Explicação***: O sistema deverá fornecer evidências, suporte ou racional para o processo ou resultado do modelo. O processo refere-se às ações, ao design e ao fluxo de trabalho do sistema, enquanto o resultado refere-se ao desfecho ou à ação realizada pelo sistema. É importante ressaltar que este princípio não estabelece critérios de qualidade para a compreensão ou a exatidão dessas explicações (30) (31).
- ***Relevância***: Esse princípio é atendido se o público-alvo for capaz de compreender a explicação. Além do público, a significância da explicação varia de acordo com o contexto e as necessidades do usuário, devendo ser adaptada ao seu nível de conhecimento e interesse no tema (30).
- ***Precisão***: A veracidade é um requisito fundamental para as explicações, que devem ser precisas de acordo com o nível de detalhe exigido pelo público. Como o nível de detalhe varia com a especialização do público, as métricas de precisão da explicação devem ser flexíveis (30).
- ***Limite de Conhecimento***: O princípio dos limites do conhecimento impõe que os sistemas sejam capazes de identificar as situações nas quais suas respostas são imprecisas, tendenciosas ou ultrapassam o escopo do modelo. Ao prevenir resultados indesejáveis, esse princípio contribui para a robustez e a confiabilidade dos sistemas (31).

Resumindo, um sistema que fornece uma explicação, mas não é compreensível, não é preciso ou está fora dos limites do conhecimento, tem valor reduzido.

Tendo em vista esses pontos, o XAI pode ser definido como: *Dada uma audiência, uma inteligência artificial explicável é aquela que produz explicações relevantes, precisas e limitadas ao escopo da aplicação para tornar seu funcionamento claro ou fácil de entender* (16) (30).

A necessidade de transparência nos sistemas tem impulsionado a pesquisa em XAI conforme demonstram os dados de uma revisão da literatura que consolidou evidências de 91 publicações no período de 2018 a outubro de 2022. O gráfico da **Figura 6** evidencia a concentração de estudos em XAI no domínio da saúde, que percentualmente possui quase cinco vezes mais estudos que finanças que ocupa a segunda posição no ranqueamento, refletindo a busca por modelos mais confiáveis e compreensíveis nesse setor (31).

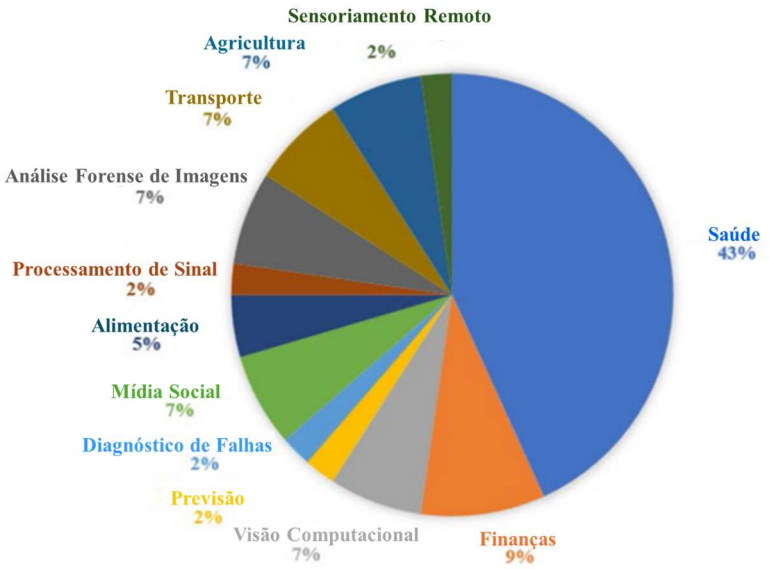


Figura 6: Representação gráfica do percentual de aplicação de XAI em diferentes setores (31)

Alguns exemplos de métodos de explicabilidade e interpretabilidade são (14):

- **Explicabilidade: Grad-CAM** (*Gradient-weighted Class Activation Mapping* ou Mapeamento de Ativação de Classe Ponderado por Gradiente*)*, Integrated Gradients, *Guided BackProp* ou Retropropagação Guiada, **LIME** (*Local Interpretable Model-agnostic Explanation* ou Explicação Local Interpretável Independente de Modelo).
- **Interpretabilidade: SHAP** (*Shapely Additive Explanations* ou Explicações Aditivas de Shapley), **TCAV** (*Testing with Concept Activations Vectors* ou Testando com Vetores de Ativação de Conceitos), *Concept Bottleneck Models* ou Modelos de Gargalos de Conceitos, **ProtoPNet** (*Prototypical Part Network* ou Rede de Partes Prototípicas).

3. MATERIAIS E MÉTODOS

A elaboração deste protocolo baseia-se na diretriz **PRISMA-P**. Esta diretriz foi desenvolvida para auxiliar na construção de protocolos de revisões sistemáticas adotando como base o processo proposto pela rede **EQUATOR** (*Enhancing the Quality and Transparency of Health Research* ou Aprimorando a Qualidade e a Transparência da Pesquisa em Saúde) (32).

O PRISMA-P fornece uma lista de 17 itens numerados (26, incluindo subitens). Os itens estão categorizados em três seções principais: informações administrativas, introdução e métodos. A finalidade dessa diretriz é assegurar a transparência e a qualidade do processo de revisão.

Foi amplamente derivada da lista de verificação do **PRISMA** (*Preferred Reporting Items for Systematic Reviews and Meta-Analyses* ou Itens Preferidos para Relato de Revisões Sistemáticas e Meta-Análises**)** e dos itens de registro do **PROSPERO** (*International Prospective Register of Systematic Reviews* ou Registro Internacional Prospectivo de Revisões Sistemáticas). Essa estratégia contribuiu tanto para consistência entre o protocolo planejado (PRISMA-P) e o relatório final da revisão quanto com o processo de registro do protocolo e da revisão.

A adoção desse guia garante que a justificativa, os métodos de busca, a seleção de estudos, a extração de dados e a análise estatística sejam explicitados antes do início da revisão, promovendo assim a rigorosidade metodológica.

Para seleção das métricas de prescrição e relato de dose serão utilizados os relatórios 83 e 91 da **ICRU** (*International Commission on Radiation Units and Measurements* ou Comissão Internacional de Unidades e Medidas de Radiação).

No que se refere tanto às métricas para segmentação quanto aos critérios para avaliação da implementação na prática clínica, na falta de uma padronização, foi realizada uma vasta busca na literatura para embasar a escolha dos parâmetros mais apropriados. O racional da decisão é compartilhado na seção de resultados.

4. RESULTADOS

4.1. DEFINIÇÃO DE CRITÉRIOS E FERRAMENTAS COM BASE NAS EVIDÊNCIAS E ANÁLISES

4.1.1. SELEÇÃO DOS SÍTIOS TUMORAIS

Com o objetivo de fomentar avanços em pesquisa e desenvolvimento de áreas que impactam a produtividade dos serviços de radioterapia, apresenta-se a seguir uma breve explanação das motivações que direcionaram a escolha dos cânceres de cabeça e pescoço, mama e próstata.

Casos de *cabeça e pescoço* frequentemente envolvem mais de 30 estruturas sensíveis, cada uma com diversas restrições dosimétricas a serem consideradas. Uma série de concessões são demandadas no planejamento de cada paciente sendo difícil assegurar que a distribuição da dose de radiação será uniforme em todas as áreas alvo, enquanto respeita-se as restrições dosimétricas das estruturas sensíveis ao redor. A garantia da homogeneização só pode ser avaliada e ajustada posteriormente, após várias iterações no planejamento. Diante disso, o processo de planejamento requer a gestão criteriosa de múltiplos fatores concorrentes, tornando-se *altamente dependente da expertise do profissional* e, ainda assim, enfrentando desafios para alcançar de forma consistente um nível de qualidade ideal (33).

O tratamento radioterápico para *câncer de mama* é geralmente estimado como responsável por pelo menos um terço de todo o trabalho em oncologia radioterápica. Representa, portanto, um dos principais focos de qualquer departamento. O tratamento deste tipo de câncer enfrenta desafios relacionados com a localização do tumor, o posicionamento do paciente, a proximidade de órgãos vitais (pulmão e/ou coração), a necessidade de ajustes devido ao movimento respiratório e a homogeneidade da dose (34) impactada tanto pelas variações em tamanho e forma na direção craniocaudal quanto pela heterogeneidade tecidual. Sem abordar os prós e os contras das múltiplas opções de tratamento.

Já o *câncer de próstata,* está entre os tipos mais prevalentes de câncer no mundo. A radioterapia é adotada como primeira linha de tratamento para um contingente

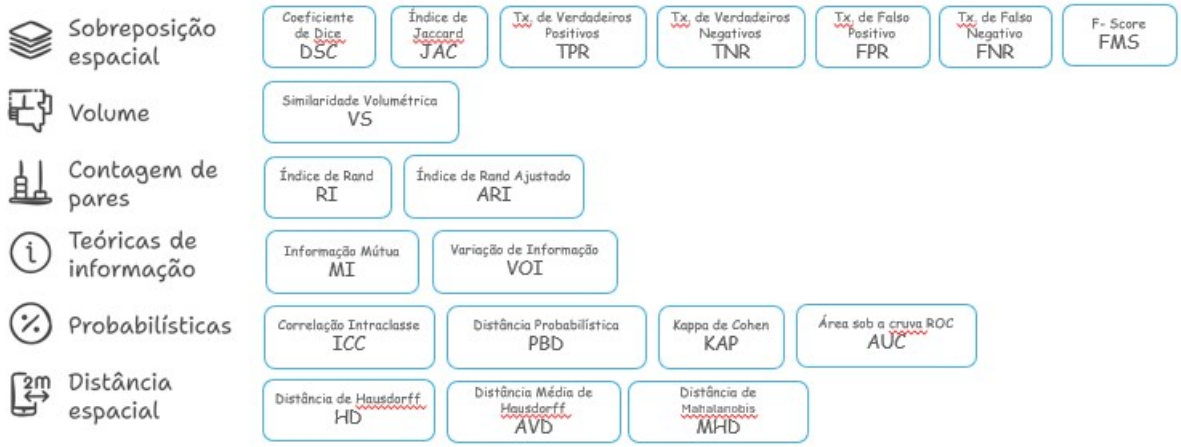
significativo de homens com esse diagnóstico, variando entre 30% e 45% dos casos, dependendo da idade (35).

4.1.2. SELEÇÃO DAS MÉTRICAS PARA AVALIAÇÃO DE SEGMENTAÇÃO

De acordo com a natureza e a definição, as métricas utilizadas para segmentação de imagem podem ser agrupadas em seis grupos (11),(36): sobreposição espacial, volume, contagem de pares, teóricas de informação, probabilística e distância espacial.

Na **Figura 7** (11), estão listados exemplos de algumas métricas obedecendo a classificação apresentada.

Figura 7: Tipos de métricas de segmentação de imagem conforme classificações (11)



Diante de tantas opções, foi necessário buscar critérios para orientar a escolha da métrica mais apropriada. Um dos materiais de apoio foi uma recomendação publicada na revista científica "*BioMed Central Medical Imaging*" cujos detalhes são compartilhados na **Tabela 1** (11) logo abaixo.

Tabela 1: Sumário dos critérios para seleção de métricas de segmentação de imagem médica (11)

	DSC	JAC	TPR	TNR	FPR	FNR	FMS	VS	GCE	RI	ARI	MI	VOI	ICC	PBD	KAP	AUC	HD	AVD	MHD
Pontos fora do padrão	✓	✓					✓	✓				✓	✓			✓	✓	X	✓	✓
Segmentação pequena	X	X	X	X	X	X	X			X	X	X	X			X	X	✓	✓	✓
Limite Complexo								X										✓	✓	X
Baixas Densidades	X	X	X	X	X	X	X	X	X	X	X	X	X	X	X	X	X	✓	✓	✓
Bx. Qual. de Segmentação								X										✓	✓	✓
Contorno Importante								X										✓	✓	X
Alinhamento Importante								X												
Sensibilidade Importante			✓									✓								
Volume Importante								✓												
Forma e Alinhamento	X	X	X	X	X	X	X	X	X	X	X	X	X	X	X	X	X	X	X	✓

Resumo diretrizes seleção de métricas. O sinal de verificação (✓) indica que a métrica é recomendada, uma célula com um (X) indica que não é recomendada, e uma célula vazia indica neutralidade em relação à recomendação.

A tabela apresenta uma comparação de diferentes métricas utilizadas para avaliar o desempenho da segmentação com base nos critérios detalhados a seguir:

Pontos fora do padrão (*outliers*) -> As segmentações automáticas podem apresentar regiões relativamente pequenas na forma de voxels (como “pixels”, mas em 3D) fora do segmento. Quando essas pequenas variações não causam prejuízos, métricas sensíveis a “*outliers”* devem ser evitadas.

Segmentação pequena -> é quando um segmento é significativamente menor que o fundo. Ou seja, quando pelo menos uma das dimensões (altura, largura ou profundidade) é muito menor do que a dimensão total da grade onde a imagem está definida. Grade é como um sistema de coordenadas que define o espaço onde a imagem é representada. Nesses casos, pequenas diferenças podem gerar resultados distorcidos. Por isso, é mais adequado utilizar métricas de distâncias em detrimento de métricas de sobreposição ou volume.

Limite complexo -> isso ocorre quando a área que está sendo avaliada tem formas irregulares ou bordas muito complexas. Por exemplo, com muitos recortes ou ondulações. Para essas aplicações, métricas sensíveis às posições dos pontos são mais adequadas.

Baixa densidade -> alguns algoritmos produzem segmentações que têm boa qualidade em termos de contorno e alinhamento, porém os segmentos não são sólidos devido a inúmeros pequenos buracos. Para esses casos, métricas que penalizam a baixa densidade devem ser evitadas. Métricas baseadas em distância são boas opções.

Baixa qualidade de segmentação -> quando a segmentação tem pouca sobreposição com a segmentação de referência. Nesses casos, as métricas baseadas em distâncias são recomendadas por serem mais capazes de detectar os desvios.

Contorno importante -> está relacionada a exatidão na delimitação de borda. Nesses casos, métricas sensíveis a posição dos pontos são as mais adequadas.

Alinhamento importante -> quando o objetivo principal é garantir que o segmento esteja na posição correta mesmo que as bordas não estejam perfeitamente precisas.

Nesses casos, as métricas de volume são desaconselhadas. Isto porque se o segmento estiver fora da posição correta, mas com o mesmo tamanho do segmento de referência, a métrica baseada em volume pode indicar um bom desempenho, mesmo que o alinhamento esteja inadequado.

Sensibilidade importante -> em alguns casos, um requisito importante é incluir tudo que deveria estar presente no segmento verdadeiro, mesmo que isso signifique incluir algumas partes que não deveriam estar lá. A delimitação precisa das bordas não é relevante. Nesses casos, devem ser priorizadas métricas que priorizem a sensibilidade (verdadeiros positivos).

Volume importante -> quando a prioridade é garantir que o tamanho total da região segmentada seja o mais próximo possível do tamanho real, mesmo que as bordas ou alinhamento não sejam perfeitos. Nesses casos, a métrica de similaridade volumétrica é recomendada.

Apesar dos elementos apresentados, dúvidas, quanto à seleção das métricas mais apropriadas, persistiam. Sendo assim, com base num aprofundamento investigativo considerando as referências bibliográficas de estudos na área de segmentação de tumores, foram selecionadas duas métricas para este estudo: **DSC** (Dice Similarity Coefficient ou Coeficiente de Similaridade de Dice) e **AVD** (Average Hausdorff Distance ou Distância Média de Hausdorff).

O DSC é amplamente utilizado em MIS (*Medical Image Segmentation* ou Segmentação de Imagens Médicas) por sua capacidade de desconsiderar o fundo e focar no objeto (37), o que é especialmente relevante em cenários onde o objeto segmentado é frequentemente pequeno em comparação ao fundo (38). Essa métrica quantifica a sobreposição entre a segmentação de referência e a segmentação gerada pela máquina, com valores variando de zero (nenhuma sobreposição) a um (sobreposição perfeita). É frequentemente recomendada para avaliar mudanças no volume tumoral, priorizando diferenças geométricas em detrimento da precisão na delimitação dos contornos.

Entretanto, como a precisão na delimitação dos contornos é igualmente importante, especialmente em tarefas de segmentação complexas, o HD (*Hausdorff Distance* ou

Distânica de Hausdorff) é amplamente utilizado como métrica complementar, pois avalia as discrepâncias de distância entre contornos (36). No entanto, devido à sua sensibilidade a "outliers" (pequenas segmentações adicionais fora do objeto principal), recomenda-se o uso de uma variante, o AVD para reduzir esse impacto. Os valores variando de 0 (ideal) a números positivos mais altos, dependendo da unidade da métrica de distância empregada (37).

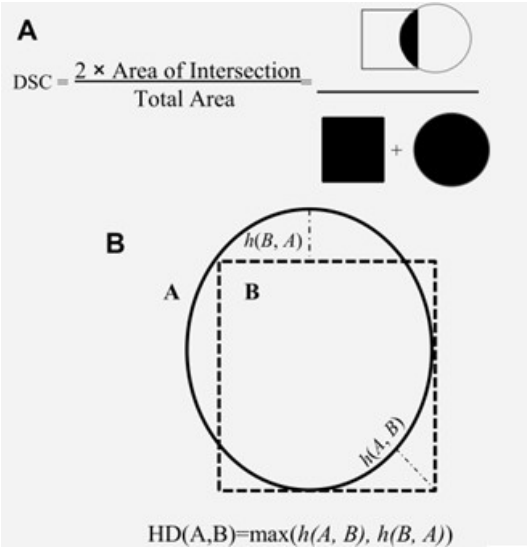


Figura 8: Ilustração esquemática de métricas de segmentação de performance (38).

Ilustração esquemática de métricas de segmentação de performance. Definição esquemática de (A) DSC e (B) HD. No esquema “B”, a segmentação de referência é a imagem A e a segmentação da máquina é a imagem B. (38)

4.1.3. SELEÇÃO DAS MÉTRICAS PARA AVALIAÇÃO DE DISTRIBUIÇÃO DE DOSE NOS VOLUMES ALVOS E EM ÓRGÃOS DE RISCO

A avaliação da distribuição de dose calculada é um elemento essencial na análise do plano de radioterapia. Contudo, esse processo é complexo devido a uma série de fatores como: protocolos, requisitos locais e práticas históricas (39). Além disso, a multiplicidade de diretrizes clínicas que variam entre diferentes regiões contribui para essa complexidade. Exemplos dessas diretrizes incluem as publicadas pelas seguintes organizações de especialistas: **AOCR** (*Asian Oceanian Congress of Radiology* ou Congresso de Radiologia da Ásia e Oceania)**, AIRO** (*Associazione Italiana di Radioterapia ed Oncologia Clinica* ou Associação Italiana de Radioterapia e Oncologia Clínica), **ASTRO** (*American Society for Radiation Oncology* ou Sociedade Americana de Oncologia por Radiação), **DAHANCA** (*Danish Head and Neck Cancer*

Group ou Grupo Dinamarquês de Câncer de Cabeça e Pescoço), **DEGRO** (*Deutsche Gesellschaft für Radioonkologie* ou Sociedade Alemã de Radioterapia e Oncologia), **ESTRO** (*European Society for Radiotherapy and Oncology* ou Sociedade Europeia de Radioterapia e Oncologia), **GORTEC** (*Groupe Oncologie Radiothérapie Tête et Cou* ou Grupo de Oncologia e Radioterapia de Cabeça e Pescoço), **RTOG** (*Radiation Therapy Oncology Group* ou Grupo de Oncologia de Terapia por Radiação), entre outras.

A variabilidade na prescrição e na administração de doses de radiação frequente nas práticas clínicas (12) dificulta a comparação de resultados e a avaliação da qualidade dos tratamentos. Padrões internacionais de prescrição de dose podem contribuir para uma uniformidade facilitando comparações intra e interinstitucionais das práticas de tratamento e seus impactos nos desfechos clínicos (40). Sob essa perspectiva, as diretrizes da ICRU estabelecem uma abordagem padronizada para delineamento de volumes-alvo e definição de parâmetros para prescrição e relato de dose. Portanto, a adesão a esses padrões deveria ser uma boa prática não limitada apenas aos ensaios clínicos (41).

Tendo em vista a complexidade da definição de métricas, para fins deste projeto, foram adotadas as diretrizes da ICRU como referência, visto que, em teoria, deveria ser a base de fundamentação técnica em radiação. Devido a escolha das áreas a serem tratadas nesse trabalho (mama, cabeça e pescoço, e próstata), foram utilizadas em particular as publicações ICRU 83 e 91. A primeira publicada em 2010 se concentra em recomendações para prescrição, registro e relato de técnicas especiais em radioterapia com feixe externo de fótons, como a **IMRT** (*Intensity-Modulated Radiation Therapy* ou Terapia de Radiação com Intensidade Modulada) para grandes volumes de tratamento. Já a segunda, publicada em 2017 é direcionada para **SRS** (*Stereotactic Radiosurgery* ou Radiocirurgia Estereotática), **SRT** (*Stereotactic Radiotherapy* ou Radioterapia Estereotática), e **SBRT** (*Stereotactic Body Radiotherapy* ou Radiocirurgia Corporal Estereotática) onde temos campos pequenos, com altas doses e extremamente hipofracionadas (42).

Em continuidade a este tópico, apresentaremos as métricas estabelecidas pela ICRU83 e ICRU91:

- *Relatório ICRU 83 (relatório de nível 2)* (43) -> **DVH**s (Dose-Volume Histogram ou Histograma Dose-Volume); **CTV** (Clinical Target Volume ou Volume Alvo Clínico) e **PTV** (Planned Target Volume ou Volume Alvo Planejado) -> D50%, Dmédia, D2%, D98%; **OAR** (Organs at Risk ou Órgãos de Risco) e **PRV** (Planning Risk Volume ou Volume de Risco Planejado) -> Dmédia, D2% e V_D;

- *Relatório ICRU 91 (relatório de nível 2)* (44)-> DVHs; CTV e PTV -> D50%, Dmédia, Dpróx.min, Dpróx.-max; OAR/PRV -> Dmédia, D2% ou D35mm3, V_D, **HI** (Homogeneity Index ou índice de homogeneidade), **CI** (Conformity Index ou índice de conformidade) e **GI** (Gradient Index ou índice de gradiente).

Onde:

D50% -> representa a dose recebida por 50% do volume de uma estrutura;

Dmédia -> dose média para a parte central da área alvo. Para PTV, a dose média absorvida e a dose mediana absorvida (D50%) são normalmente muito próximas;

D98% -> também conhecida como Dprox.mín. (100%), é a dose mínima recebida por 98% do volume planejado. Ou seja, é a dose de radiação que quase todo o volume alvo recebe garantindo assim que a maior parte do tumor seja adequadamente tratada;

D2% -> também conhecida como Dprox.máx.(0%), refere-se à dose máxima recebida por 2% do volume planejado. Indica a dose mais alta entregue a uma pequena fração do volume ou órgão avaliado. Por isso, é utilizada para avaliar o risco de sobredosagem em regiões específicas, o que poderia causar toxicidade ou efeitos colaterais indesejados. Se outro descritor de dose-volume for considerado clinicamente relevante (por exemplo, D1% para câncer de nasofaringe de acordo com o protocolo 0615 do RTOG), recomenda-se relatar também D2%. Esse valor é simples de obter e agrega consistência ao relatório;

V_D-> indica o volume que está recebendo pelo menos a dose D. Essa dose "D" indica um nível de dose acima do qual há uma alta probabilidade de causar complicações significativas.

Na ICRU 91, temos algumas mudanças no conceito de Dprox.min e Dprox.max. além da inserção de outras métricas para relatório de nível 2 (avançado - técnicas de ponta) detalhadas a seguir (44):

Dprox.min -> Para o PTV onde V é maior ou igual a 2 cm³, o volume próximo do mínimo representa 98% do PTV, conforme recomendado no Relatório ICRU 83 (D98 %). Para o PTV onde V é menor que 2 cm³, o próx.-min. é um volume absoluto de 35 mm³, e nesse caso, $D_{V\text{-}35mm3}$ é relatado;

Dprox.max. -> Para o PTV onde V é maior ou igual a 2 cm³, o volume próximo do máximo representa 2% do PTV, conforme recomendado no Relatório 83 da ICRU (D2 %). Para o PTV onde V é menor que 2 cm³, o prox.-max. é um volume absoluto de 35 mm³, e nesse caso, D_{35mm^3} é relatado;

Índice de homogeneidade -> caracteriza a uniformidade da distribuição da dose dentro do volume alvo a fim de garantir que a dose prescrita seja distribuída de maneira homogênea minimizando grandes variações de dose entre diferentes partes da região de interesse. Tem uma importante relevância clínica no tratamento de metástases cerebrais. Com base na ICRU83 para relatório nível 3 (pesquisa e desenvolvimento inovadores), temos a seguinte definição:

$$HI = \frac{D2\% - D98\%}{D50\%}$$

Índice de conformidade -> a conformidade da dose caracteriza o grau em que a região de alta dose se conforma ao volume-alvo, geralmente o PTV, garantindo que a dose prescrita cubra adequadamente a região de interesse limitando a exposição desnecessária de tecidos adjacentes. Ou seja, está relacionada com a precisão que a dose se ajusta ao volume alvo evitando irradiar áreas saudáveis. O índice é definido como o quociente entre o volume-

alvo (TV) e o volume da isodose de prescrição (PIV), com o volume-alvo dentro do volume da isodose de prescrição (TV_{PIV}).

$$CI = \frac{TVxPIV}{TV^2{}_{PIV}}$$

Índice de Gradiente -> é uma métrica para avaliar a conformidade entre a dose prescrita e a forma do PTV. Para radiocirurgia no cérebro, é recomendado considerar essa métrica definida como:

$$GI = \frac{PIV_{half}}{PIV}$$

Onde PIV_{half} representa o volume de isodose prescrito na metade da isodose prescrita (por exemplo, a 25%) e PIV, o volume total de isodose prescrito (por exemplo, a 50%).

4.1.4. FERRAMENTAS PARA AVALIAÇÃO DE ESTUDOS DE IA

A IA é uma área disruptiva da ciência da computação com múltiplas pesquisas de caso de uso sendo realizadas em todo setor de saúde (25). Na radioterapia, em específico, existe a expectativa de que sua aplicação possa contribuir para melhorar a precisão, eficiência e qualidade geral do tratamento dos pacientes oncológicos (21).

Para garantir a robustez, qualidade, ética e clareza das publicações nessa área, desde 2020 se observa uma intensificação da publicação de diretrizes específicas orientadas a padronizar a comunicação, evitar falhas metodológicas e assegurar a validade dos resultados conforme ilustrado na **Figura 9** (9).

•**Diretrizes pré-clínicas** regulam estudos teóricos que não envolvem dados de desfecho clínico, mas que podem, potencialmente, envolver dados de pacientes de forma retrospectiva.
•**Diretrizes translacionais** focam em estudos retrospectivos ou prospectivos observacionais que utilizam dados de pacientes com possíveis implicações clínicas.
•**Diretrizes clínicas** regulam estudos intervencionais em um ambiente clínico.
A abrangência das diretrizes é classificada como **geral** ou **específica para determinado assunto**, dependendo das áreas de pesquisa alvo mencionadas na diretriz.
Quanto ao processo de consenso:
•**Diretrizes abrangentes** baseiam-se em uma abordagem estruturada, consensual e metódica, desenvolvida com a participação de múltiplos especialistas e partes interessadas relevantes, com detalhes sobre o protocolo exato.
•**Diretrizes colaborativas** são presumidamente desenvolvidas utilizando um procedimento formal de consenso com múltiplos especialistas, mas não fornecem detalhes sobre o protocolo exato ou a estrutura metodológica.
•**Diretrizes lideradas por especialistas** não são desenvolvidas por meio de um procedimento baseado em consenso, não envolvem partes interessadas relevantes ou não descrevem claramente o procedimento de desenvolvimento.

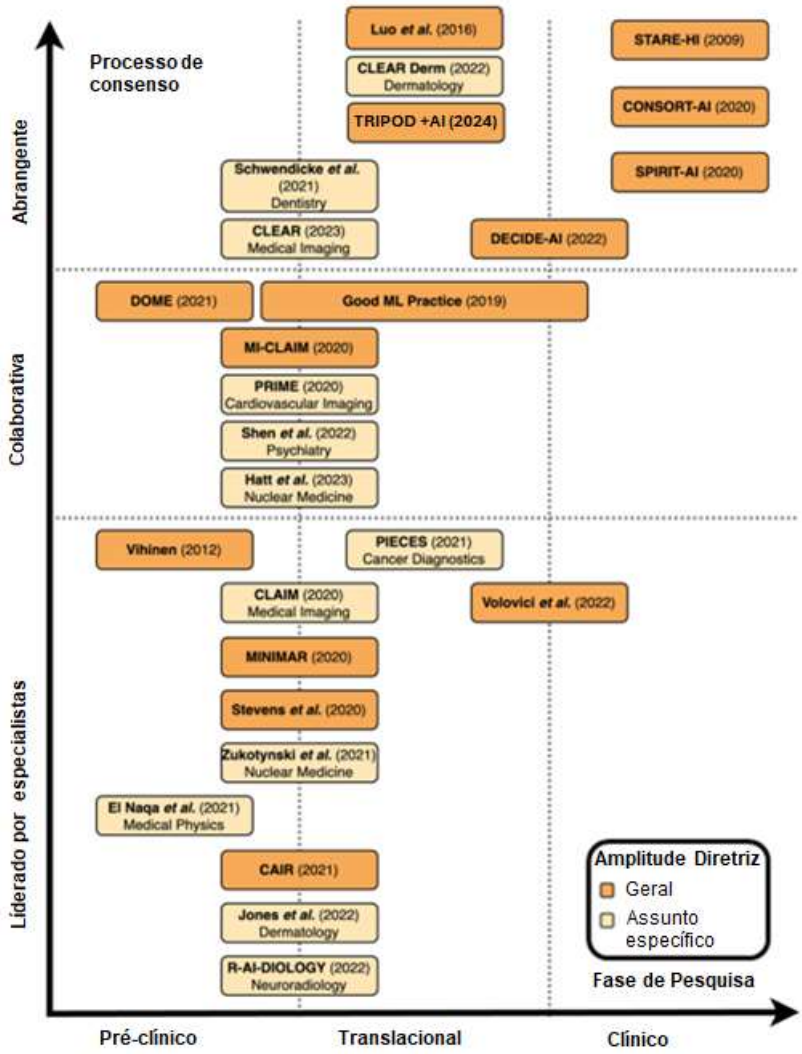


Figura 9: Diretrizes de relato existentes sobre IA na medicina por fase de pesquisa e nível de consenso (9)

Quando o tema é o planejamento radioterápico, foco deste trabalho, a precisão na identificação e delimitação dos volumes-alvo e dos órgãos de risco é crucial influenciando diretamente a qualidade do tratamento, a segurança do paciente e o desfecho clínico (45)(46).

De acordo com uma revisão sistemática publicada em 2020 (46), 90% das diretrizes de contorno avaliadas recomendaram a delimitação em exames de TC e 10% ofereceram recomendações exclusivas para contornos baseados em ressonância magnética (RM). Portanto, é de fundamental importância que sejam consideradas ferramentas orientadas à avaliação de imagens médicas.

Conforme já mencionado na seção reservada ao objetivo deste trabalho, busca-se aprofundar a análise dos algoritmos de IA ultrapassando a mera perspectiva de performance incluindo aspectos como dados (diversidade, qualidade e confiabilidade), treinamento do modelo, rastreabilidade, robustez, reprodutibilidade e explicabilidade. Sendo assim, para comparar as múltiplas ferramentas disponíveis nestas diferentes dimensões, fez-se uso da análise de concordância entre elas conforme ilustrado na **Figura 10** (9). Nesta imagem, os temas abordados são categorizados em cinco grupos: *justificativa clínica, dados, treinamento e validação de modelos, avaliação crítica, ética e reprodutibilidade*. A intensidade da cor em cada célula da figura indica a importância atribuída a cada tema em cada grupo, permitindo identificar as áreas de maior e menor ênfase. Quanto mais escuro maior a ênfase.

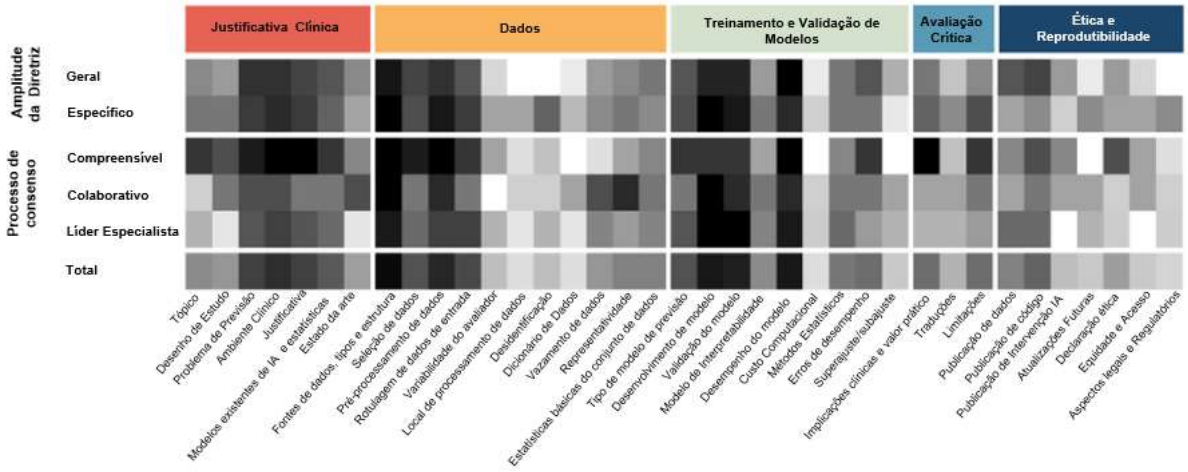


Figura 10: Concordância das diretrizes de relato voltadas a IA no setor de saúde (9)

Analisando as informações disponibilizadas, é possível observar que temas como interpretabilidade/explicabilidade (comutação frequente relatada anteriormente), robustez e rastreabilidade não são abordados ou quando abordados a ênfase é baixa. Diante da identificação dessa lacuna, decidiu-se adotar como referência o **FUTURE-AI** (6) (*Fairness, Universality, Traceability, Usability, Robustness and Explainability* ou Imparcialidade, Universalidade, Rastreabilidade, Usabilidade, Robustez e Explicabilidade). Embora não seja uma ferramenta formal de avaliação, é uma diretriz direcionada a desenvolvedores que aborda esses aspectos de forma mais abrangente e consistente.

Considerando o exposto, a seguir são apresentadas as ferramentas selecionadas para direcionar a extração de dados considerando as delimitações do trabalho.

- ✓ **CLAIM** (*Checklist for Artificial Intelligence in Medical Imaging*) (47) -> Lista de Verificação para Inteligência Artificial em Imagem Médica;
- ✓ **MINIMAR** (*MINimum Information for Medical AI Reporting*) (48) -> Informação Mínima para Relatórios de IA Médica
- ✓ **MI-CLAIM** (*Minimum Information about Clinical Artificial Intelligence Modeling*) (15) -> Informação Mínima sobre Modelagem de Inteligência Artificial Clínica
- ✓ **FUTURE-AI** (14) -> Como já mencionado, não é uma ferramenta de avaliação. Por esse motivo, não consta na relação de ferramentas da **Figura 9** tampouco está inclusa na avaliação exposta na **Figura 10**. Publicado em 2021, consiste em princípios orientadores para desenvolvedores os quais foram consolidados a partir da experiência acumulada, no consenso e nas melhores práticas de cinco grandes projetos europeus sobre IA em imagem médica. Ao longo do documento são discutidos os princípios que dão origem ao acrônimo, FUTURE (imparcialidade, universalidade, rastreabilidade, usabilidade, robustez e explicabilidade), apresentando alternativas para mitigá-los. Ao final, é proposta uma lista de verificação para orientação de desenvolvedores.

4.1.5. FERRAMENTA PARA AVALIAÇÃO DE RISCO DE VIÉS

Para *modelos de predição que envolvem prognósticos e diagnósticos*, já existem diretrizes estabelecidas para avaliar tanto a qualidade metodológica quanto o risco de viés. Nessa ordem, podemos citar, como exemplos, o **TRIPOD** (*Transparent Reporting of a Multivariable Prediction Model for Individual Prognosis or Diagnosis* ou Relato Transparente de um Modelo de Predição Multivariável para Prognóstico ou Diagnóstico Individual), e o **PROBAST** (*Prediction Model Risk of Bias Assessment Tool* ou Ferramenta de Avaliação para Risco de Viés em Modelos de Predição). Mais recente, em janeiro de 2024, foi publicado o **TRIPOD+AI** voltado para avaliação da qualidade metodológica de modelos que utilizam técnicas de IA (49).

No entanto, essas ferramentas não podem ser aplicadas diretamente para avaliar estudos sobre o desempenho da IA em radioterapia, particularmente na segmentação automática e no planejamento de tratamento, sem modificação ou esclarecimento.

Isso ocorre porque certos termos, normalmente usados no contexto de prognóstico individual ou predição diagnóstica, podem não ser totalmente claros na avaliação de desempenho da IA em oncologia por radiação (50).

Portanto, para avaliar o risco de viés, foi empregada uma versão adaptada da lista de verificação PROBAST, que foi desenvolvida através de um processo Delphi para estudos de radioterapia baseados em IA (50). Na **Tabela 2**, a relação completa dos itens revisados.

Tabela 2: Checklist PROBAST Revisado (50)

Ferramenta PROBAST revisada				
Domínio	**Número**	**Item da Lista**	**Item Revisado**	**Esclarec. Adicionais**
1. Participantes	A (risco de viés)	Descreva as fontes de dados e os critérios para a seleção dos participantes.	Mantido idêntico	
	1.1	Foram usadas fontes de dados apropriadas, por exemplo, dados de coorte, ECR (Ensaios Clínicos Randomizados) ou estudo de caso-controle aninhado?	Mantido idêntico	
	1.2	Todas as inclusões e exclusões de participantes foram apropriadas?	Mantido idêntico	
	B (Aplicabilidade)	Descreve os participantes incluídos, ambiente/contexto e as datas	Descreve os participantes incluídos, ambiente/cenário e datas **assim como uma descrição das características do plano, se aplicável**	

Domínio	Número	Item da Lista	Item Revisado	Esclarec. Adicionais
2. Preditores (nome de domínio revisado: parâmetros de entrada da IA)	A (risco de viés)	Liste e descreva os preditores incluídos no modelo final, por exemplo, definição e tempo de avaliação	Liste e descreva os **parâmetros de entrada do modelo de IA** incluídos no modelo final, por exemplo, definição e tempo de avaliação, **modalidades de imagem utilizadas para o planejamento e os respectivos parâmetros como espessura de corte de TC ou tamanho de voxel de dose, etc.**	
	2.1	Os preditores foram definidos e avaliados de maneira semelhante para todos os participantes?"	Os **parâmetros de entrada do modelo de IA** foram definidos e avaliados **da mesma forma** para todos os participantes?	
	2.2	A avaliação dos preditores foram realizadas sem conhecimento dos dados de desfecho?	As avaliações dos **parâmetros de entrada** foram realizadas sem o conhecimento dos **dados de saída do modelo de IA**?	
	2.3	Todos os preditores estão disponíveis no momento em que o modelo se destina a ser utilizado?	Todos os **parâmetros de entrada** estão disponíveis no momento em que o modelo se destina a ser utilizado?	
	B (aplicabilidade)	Preocupação de que a definição, avaliação ou o momento dos preditores no modelo não correspondam à questão da revisão	Não aplicável	
3. . Desfecho (nome do domínio revisado: saída do modelo de IA)	A (risco de viés)	Descreve o desfecho, como ele foi definido e determinado, e o intervalo de tempo entre a avaliação dos preditores e a determinação do desfecho	Descreva a **saída do modelo de IA** e como ela foi definida	
	3.1	O desfecho foi determinado de forma apropriada	O **resultado do modelo de IA** foi determinado de forma apropriada?	
	3.2	Uma definição de desfecho pré-especificada ou padrão foi utilizada?	Uma definição de **saída de modelo de IA** pré-especificada ou padrão foi utilizada?	
	3.3	Os preditores foram excluídos da definição do desfecho?	Os **parâmetros de entrada** foram excluídos da definição da saída do modelo?	
	3.4	O desfecho foi definido e determinado de maneira semelhante para todos os participantes?	O **resultado do modelo de IA** foi definido e determinado de maneira semelhante para todos os participantes?	
	3.5	O desfecho foi determinado sem conhecimento das informações dos preditores?	O **resultado do modelo de IA** foi determinado sem conhecimento das informações dos **parâmetros de entrada**?	

Domínio	Número	Item da Lista	Item Revisado	Esclarec. Adicionais
	3.6	O intervalo de tempo entre a avaliação dos preditores e a determinação do desfecho foi apropriado?	Não aplicável	
	B (aplicabilidade)	Em que momento o desfecho foi determinado: Se um desfecho composto foi utilizado, descreva a frequência relativa/distribuição de cada desfecho contribuinte	Não aplicável	
4. Análise	A	Descreva o número de participantes, o número de parâmetros de entrada selecionados e a saída do modelo	Mantido idêntico	
	B	Descreva como o modelo foi desenvolvido (por exemplo, em relação à técnica de modelagem (por exemplo, modelagem de sobrevivência ou logística), seleção de preditores e definição de grupos de risco).	Descreva como o modelo foi desenvolvido (por exemplo, em relação à técnica de modelagem (tipo de modelo) e seleção de **parâmetros de entrada**)	
	C	Descreva se e como o modelo foi validado, seja internamente (por exemplo, autoajuste, validação cruzada, amostra aleatória dividida) ou externamente (por exemplo, validação temporal, validação geográfica, diferentes configurações, diferentes tipos de participantes).	Mantido idêntico	
	D	Descreva as métricas de performance do modelo, por exemplo: recalibração, discriminação, reclassificação, benefício líquido e se foram ajustadas por otimismo.	Mantido idêntico	
	E	Descreva quaisquer participantes que foram excluídos da análise	Mantido idêntico	
	F	Descreva os dados ausentes sobre preditores e desfechos, bem como os métodos utilizados para lidar com dados ausentes.	Descreva os dados ausentes sobre os **parâmetros de entrada** e a **saída do modelo de IA**, bem como os métodos utilizados para lidar com dados ausentes.	
	4.1	Havia um número razoável de participantes no desfecho?	Havia um número razoável de participantes no **resultado do modelo de IA**?	
	4.2	Os preditores contínuos e categóricos foram gerenciados de forma apropriada?	Os **parâmetros de entrada** contínuos e categóricos foram tratados de forma apropriada?	
	4.3	Todos os participantes inscritos foram incluídos na análise?	Mantido idêntico	
	4.4	Os participantes com dados ausentes foram tratados de forma adequada?	Mantido idêntico	

Domínio	Número	Item da Lista	Item Revisado	Esclarec. Adicionais
	4.5	A seleção de preditores com base em análise univariada foi evitada?	Não aplicável	
	4.6	As complexidades nos dados (por exemplo, censura, riscos concorrentes, amostragem de controles) foram consideradas de forma apropriada?	As complexidades nos dados foram consideradas de forma apropriada?	O texto entre parênteses foi omitido, pois não tornava a pergunta mais clara
	4.7	As medidas de performance relevantes para o modelo foram avaliadas adequadamente?	Mantido idêntico	
	4.8	O sobreajuste do modelo e o otimismo no desempenho do modelo foram levados em consideração	Mantido idêntico	
	4.9	Os preditores e seus pesos atribuídos no modelo final correspondem aos resultados da análise multivariada?	Não aplicável	

4.1.6. FERRAMENTA PARA AVALIAÇÃO DA QUALIDADE DO ESTUDO E DA SÍNTESE DE EVIDÊNCIA

Antes de abordar as ferramentas, é relevante destacar que, na radioterapia, os estudos de planejamento de tratamento são fundamentais para apoiar e facilitar a implementação de novas técnicas na prática clínica. Esses estudos utilizam simulações computacionais como uma alternativa aos experimentos laboratoriais ou clínicos convencionais, permitindo que os pesquisadores explorem e avaliem estratégias e métodos inovadores sem a necessidade imediata de ensaios clínicos (51) .

Esses estudos não envolvem manipulação direta do objeto de pesquisa no mundo real. Em vez disso, analisam dados históricos ou criam modelos computacionais que simulam o comportamento do sistema em análise. Esses modelos, também, podem ser aplicados no desenvolvimento de sistemas de predição. Nesses casos, se aproximam mais de um experimento computacional do que de uma simples pesquisa sem intervenção direta focada em correlação e padrões.

Dado o escopo, é importante assegurar a qualidade dos estudos utilizados. É preciso fornecer informações detalhadas e consistentes para garantir que outros pesquisadores possam reproduzir os resultados de modo fidedigno e verificar a validade dos resultados. Isso é essencial para a aplicação segura e eficaz dos

métodos em diferentes contextos (51). No entanto, considerando a natureza desse tipo de estudo bem como as especificidades da área, ferramentas desenvolvidas para avaliação de estudos clínicos randomizados, não randomizado ou observacionais podem não ser adequadas para capturar todas as dimensões da pesquisa em questão.

Segundo o relatório da Comissão Global de Evidência, ainda não há uma ferramenta amplamente reconhecida para avaliar a qualidade no campo da inteligência artificial (52). No entanto, no domínio de avaliação de estudos de planejamento de tratamento, dentro dos recursos disponíveis, o **RATING** (*Radiotherapy Treatment plannINg study Guidelines* ou Estudos de Planejamento de Tratamento em Radioterapia) se apresenta como uma alternativa mesmo não sendo específica para IA (51).

Em relação à ferramenta para avaliação da síntese de evidências, a aplicação direta do **GRADE** (53) (*Grading of Recommendations Assessment, Development, and Evaluation* ou Graduação das Recomendações de Avaliação e Desenvolvimento) não é recomendável. Isso ocorre porque o ponto de partida para a avaliação da qualidade das evidências no GRADE é o delineamento do estudo. Assim, ensaios clínicos randomizados são classificados inicialmente como evidências de alta qualidade; enquanto estudos observacionais são considerados evidências de baixa qualidade desde o início. Diferentemente de outras áreas da saúde que se concentram em estudos "in vitro" ou "in vivo", a radioterapia se destaca pelo uso extensivo de simulações computacionais ("in silico") como principal ferramenta de pesquisa. Em outras palavras, o ponto de partida dos estudos é fundamentalmente diferente.

4.1.7. EXAMES DE IMAGEM

Para fins desse estudo, não se farão restrições em relação ao tipo de exame de imagens. Isto porque na segmentação de tumores a TC, a RM e o **PET-CT** (*Positron Emission Tomography - Computed Tomography* ou Tomografia por Emissão de Pósitrons – Tomografia Computadorizada) podem ser usados em conjunto para uma avaliação mais completa.

A TC e a RM são as modalidades mais amplamente utilizadas para visualizar detalhes anatômicos. Sendo a primeira mais recomendada para visualizar estruturas ósseas e

áreas com tecidos de densidades diferentes. Por apresentar um custo inferior, é encontrada com mais facilidade nos serviços de saúde.

Quanto a RM, fornece informações semelhantes às da TC. No entanto, a aquisição de imagens é mais lenta e, portanto, mais suscetível à variação de movimento. A RM tem uma resolução mais alta e permite uma melhor detecção de anomalias sutis nos tecidos moles. Por essa razão, é mais adequada para visualizar tumores que não apenas formam uma massa discreta, mas também invadem os tecidos ao seu redor como é o caso do gliomas (tumores do sistema nervoso central) e do câncer de mama (54).

Com intuito de trazer informações sobre a utilização de TC e RM em segmentação de tumores nas áreas de interesse deste estudo, são mostrados a seguir dados coletados de uma revisão sistemática publicada em 2019 (55):

- *Segmentação de tumores neurológicos*: de 145 estudos selecionados, 94% utilizaram RM, 5% utilizaram TC e 1% utilizou TC e RM;
- *Segmentação de tumores na mama*: de 20 estudos selecionados, todos utilizaram RM;
- *Segmentação de tumores abdominais*: de 87 estudos selecionados, 57% utilizaram TC, 41% utilizaram RM e 2% utilizaram TC e RM.

No que se refere ao PET-CT, é uma modalidade diagnóstica que integra a tomografia por emissão de pósitrons utilizando 18F-fluordesoxiglicose (FDG-PET) com a TC. Essa combinação permite tanto uma análise morfológica (tamanho, densidade, contorno etc.) quanto funcional (presença ou não de metabolismo glicolítico) de tumores. Suas limitações incluem: tumores com subtipos que apresentam baixa captação de glicose, casos de comprometimento cerebral secundário assim como casos que envolvam processos inflamatórios ou infecciosos (56). Sua aplicação no caso de câncer de mama, por exemplo, pode ser recomendada no planejamento radioterápico para tratar a recidiva de um câncer numa área localizada próxima ao tumor original. Já no caso de câncer cerebral, pode ser indicado para diferenciar cicatrizes pós-tratamento de doença residual ou recorrente (54).

4.2. ESTRUTURAÇÃO DO PROTOCOLO

4.2.1. PERGUNTAS A SEREM RESPONDIDAS:

Ratificando o que foi delineado na seção “Objetivo”, este trabalho tem como foco propor uma metodologia para avaliar a segmentação e o planejamento dosimétrico, considerando tanto a performance quanto a aplicabilidade clínica. Portanto, é pertinente que as questões sejam organizadas de maneira a seguir essa estrutura.

4.2.1.1. AVALIAÇÃO DESEMPENHO DO ALGORITMO:

a) Qual é a **exatidão dos algoritmos de IA na** *segmentação de tumores* em *pacientes com câncer de cabeça e pescoço, mama ou próstata submetidos à radioterapia*, considerando as **métricas DSC e AVD** tendo como *comparador os profissionais de saúde ou sistemas computacionais não baseados em IA*?

b) Qual é a **exatidão dos algoritmos de IA no** *planejamento dosimétrico* (mapa de distribuição de dose) em *pacientes com câncer de cabeça e pescoço, mama ou próstata submetidos à radioterapia*, de acordo com os **volumes e parâmetros dosimétricos** estabelecidos pelos **relatórios 83 ou 91 da ICRU** tendo como *comparador os profissionais de saúde ou sistemas computacionais não baseados em IA*?

Nota: Maiores informações sobre o racional envolvido na definição das métricas de performance, estão disponíveis nas seções “4.1.2 Seleção das métricas para avaliação de segmentação” e “4.1.3 Seleção das métricas para avaliação de distribuição de dose nos volumes alvos e em órgãos de risco”. As métricas específicas do planejamento dosimétrico estão listadas na tabela de extração de dados disponível no item “4.2.5 Itens de dados”.

4.2.1.2. AVALIAÇÃO DA VIABILIDADE DE APLICAÇÃO NA PRÁTICA CLÍNICA:

Qual é a **viabilidade de implementação clínica dos algoritmos de IA voltados a** *segmentação e/ou planejamento dosimétrico* em *pacientes com câncer de cabeça e pescoço, mama ou próstata submetidos à radioterapia*

considerando as seguintes **perspectivas: dados, confiabilidade dos dados, treinamento, rastreabilidade, robustez, reprodutibilidade e explicabilidade**?

Nota: Maiores informações sobre as razões que nortearam a escolha dessas perspectivas, estão disponíveis no item "2.3 Os desafios para incorporação na prática clínica". Se o foco for na fundamentação para garantir a cobertura desses aspectos, as informações encontram-se acessíveis no item "4.1.4 Ferramentas para avaliação de estudos de IA". Caso deseje acessar diretamente os critérios, foram disponibilizados na tabela de extração de dados no item "4.2.5 Itens de dados".

4.2.2. CRITÉRIO DE ELEGIBILIDADE:

4.2.2.1. TIPO DE ESTUDO

Para ampliar a base de evidências, recomenda-se a inclusão de quaisquer relatórios científicos que avaliem o uso da inteligência artificial para segmentação de tumor e/ou planejamento dosimétrico. Sugere-se a exclusão de comentários, editoriais, estudo de caso, artigo de opinião, artigo de correspondência e outras publicações informais. Além disso, é recomendável a exclusão de estudos relacionados a braquiterapia, IA focado na análise de biomarcadores ou em qualquer outra etapa do fluxo de trabalho da radioterapia que não abranja as duas áreas específicas de interesse.

4.2.2.2. PARTICIPANTES

Orienta-se a inclusão de indivíduos diagnosticados com câncer de cabeça e pescoço, mama e próstata que necessitem de tratamento radioterápico em qualquer estágio de estadiamento da doença, sem restrições de idade, gênero, etnia ou localização geográfica.

Nota: Uma base de pacientes mais ampla não apenas aumenta as chances de generalização dos resultados, mas também contribui para um treinamento mais robusto de modelos, reduzindo vieses e promovendo maior imparcialidade. Uma vez que, contribui para mitigar/evitar imprecisões para certos grupos populacionais. Essa abordagem favorece a equidade de acesso aos avanços científicos, garantindo que as soluções desenvolvidas sejam mais representativas e aplicáveis a diversas populações/regiões.

4.2.2.3. INTERVENÇÃO

Indica-se que sejam consideradas ferramentas que utilizem algoritmos de inteligência artificial para etapa segmentação de tumor e/ou planejamento dosimétrico associada a exames de imagem (TC, RM ou PET-CT).

Nota: Para mais detalhes sobre a ausência de restrições em relação ao tipo de imagem consultar o item "4.1.6 Exames de Imagem".

4.2.2.4. COMPARADORES

É sugerido que sejam considerados tanto profissionais de saúde quanto sistemas computacionais voltados à otimização do planejamento de radioterapia que não sejam baseados em inteligência artificial.

Nota: Além de expandir a base de estudos, a dupla comparação, com humanos e sistemas computacionais não baseados em IA, possui dois objetivos principais: (a) verificar se o algoritmo de IA consegue, no mínimo, alcançar os padrões de desempenho humano; (b) analisar se os algoritmos oferecem vantagens adicionais em relação aos sistemas existentes, como maior eficiência ou superioridade de desempenho, a fim de justificar o investimento em novas tecnologias.

4.2.2.5. PERÍODO DAS PUBLICAÇÕES

Cabe incluir publicações de qualquer data, sem restrições quanto ao período de publicação.

Nota: Embora o volume de estudos em IA tenha aumentado significativamente a partir de 2018, a inclusão de trabalhos anteriores pode contribuir para identificar tendências, lacunas de conhecimento assim como, eventuais, alterações em relação a desafios e avanços ao longo dos anos.

4.2.2.6. IDIOMA

É adequado que a busca seja realizada sem restrições de idioma.

4.2.3. FONTE DE INFORMAÇÃO E ESTRATÉGIA DE BUSCA

Recomenda-se que a estratégia de busca seja desenvolvida utilizando recursos como **MeSH** (*Medical Subject Headings* ou Cabeçalho de Assuntos Médicos) e o **EMTREE** (*EMBASE Medical Thesaurus* ou Tesauro Médico do EMBASE). Considerando o escopo da pesquisa, é sugerido o uso dos seguintes descritores: inteligência artificial, inteligência artificial explicável, radioterapia, sistema de planejamento radioterápico; câncer de mama, cabeça e pescoço e próstata e braquiterapia. Além do uso dos vocabulários controlados, é aconselhável que os descritores sejam pesquisados em campos como: títulos, resumos, palavras-chaves, tema ou tópico. A sintaxe deve ser adaptada de acordo com os recursos disponíveis nas respectivas bases de dados selecionadas para busca. Convém realizar buscas em bases de dados oficiais incluindo PUBMED, Web of Science, Cochrane Library e EMBASE, mas não se limitando apenas a estas. Além da busca nas bases oficiais, é apropriado uma busca por diretrizes práticas, publicações de agências governamentais e relatórios nas bases de literaturas cinzas, como por exemplo, as elencadas a seguir: **HTA** database (*Health Technology Assessment Database* ou Base de Dados de Avaliação de Tecnologias em Saúde), **WHO** Library Catalog (*World Health Organization Library Catalog* ou Catálogo da Biblioteca da Organização Mundial da Saúde) e **AHRQ**

(*Agency for Healthcare Research and Quality* ou Agência para Pesquisa e Qualidade em Saúde).

Nota: Tendo em vista o rápido desenvolvimento e as inovações contínuas da IA, a inclusão da literatura cinza tem como intuito assegurar que todas as evidências disponíveis sejam levadas em conta.

4.2.4. ARMAZENAMENTO DOS DADOS

4.2.4.1. GESTÃO DOS DADOS

Para o processo de triagem e seleção de estudos para leitura completa, é sugerido o uso de ferramentas como o Rayyan Software, um programa baseado na Internet que facilita a colaboração entre revisores.

Caso essa ferramenta seja utilizada, o recurso de auto resolução pode ser empregado para tratar duplicações. Essa funcionalidade é indicada para resoluções automáticas em casos com níveis de similaridade iguais ou superiores a 95%. Para casos abaixo desse limite, é altamente aconselhável a resolução manual.

Para o processo de resolução manual, com base nas perguntas a serem respondidas e nos critérios de elegibilidade, segue uma orientação de lista de palavras-chaves:

- *Palavras chaves para inclusão*: inteligência artificial, aprendizado de máquina, aprendizado profundo, planejamento de tratamento radioterápico, segmentação de tumores, dose de radiação, DVH, DSC, AVD, dose média, dose máxima, dose mínima, câncer de cabeça e pescoço, câncer de mama, câncer de próstata, radioterapia, planos de tratamento automatizados e ferramentas de explicabilidade;

- *Palavras chaves de exclusão:* comentários, editoriais, estudo de caso, artigo de opinião, artigo de correspondência, braquiterapia, análise de biomarcadores, decisões de tratamento, otimização de feixe, decomposição dose feixe, controle de qualidade, posicionamento do paciente, predição de resposta ao tratamento, predição de evento adverso.

Para o gerenciamento de referências, podem ser utilizadas ferramentas como o Mendeley.

Destaca-se a necessidade e relevância da elaboração de uma planilha para orientar a fase de extração de dados. No item “4.2.5 Itens de Dados” é disponibilizado um modelo.

Para facilitar a colaboração e o acesso às informações, é importante prever uma plataforma para compartilhamento de arquivos, como o Google Drive. Isso permitirá a centralização dos documentos e o acesso a versões atualizadas.

4.2.4.2. PROCESSO DE SELEÇÃO

Uma vez executado o processo de exclusão dos arquivos duplicados, todos os artigos restantes devem ter seus títulos e resumos revisados. Aconselha-se que essa revisão seja realizada por, no mínimo, dois avaliadores independentes. Tendo como referência a lista de palavras de inclusão e exclusão, os estudos podem ser classificados numa das seguintes categorias:

- Incluir
- Talvez
- Excluir com justificativa

Em caso de discrepâncias, recomenda-se que os revisores discutam para alcançar um consenso. Estudos classificados como "incluir" ou "talvez" devem ser avaliados integralmente por pares de revisores para confirmar sua elegibilidade. Se houver discordâncias, deve-se seguir o procedimento padronizado para resolução. É importante especificar os motivos para a exclusão dos artigos. Adicionalmente, é apropriado que todos os revisores tenham acesso às informações sobre os títulos das revistas, autores e instituições dos estudos

Nota: Destaca-se a importância da inserção das justificativa no processo de exclusão dando transparência em relação as razões pelas quais determinados estudos foram descartados. Este detalhamento, pode ser reportado no diagrama de fluxo do PRISMA aumentando a confiabilidade e a reprodutibilidade da pesquisa. Além de permitir que outros pesquisadores avaliem a pertinência das decisões tomadas.

4.2.4.3. Processo de Coleta de Dados

Concluída a triagem, os dados serão extraídos e tabulados respeitando a tabela de extração de dados pré-definida.

Para essa proposta de metodologia de avaliação, foi elaborada uma tabela de extração de dados base nas seguintes ferramentas: CLAIM, FUTURE-AI, MI-CLAIM e MINIMAR.

Em linha com as perguntas a serem respondidas, a tabela de extração sugerida inclui as seguintes perspectivas: avaliação da performance da segmentação, avaliação da performance dosimétrica (CTV, PTV, OAR e PRV), dados, confiabilidade dos dados, treinamento do modelo, rastreabilidade, robustez, reprodutibilidade e explicabilidade.

Propõem-se que a extração e tabulação seja executada em duplicata por revisores independentes.

Notas: A extração de dados pode ser realizada por um revisor com verificação de outro, para reduzir vieses e erros no processo. Nesses casos, é importante que seja declarada de forma explicita a opção pela extração individual a fim de que revisores e leitores estejam cientes da possibilidade de erros na revisão final.

Maiores detalhes em relação a seleção das ferramentas estão disponíveis na seção "4.1.4 Ferramentas para avaliação de estudos de IA".

4.2.5. Itens de Dados

A seguir, os critérios de avaliação, bem como as respectivas ferramentas utilizadas como referência.

	FONTE	PERSPECTIVA	Nr.	DESCRIÇÃO
AVALIAÇÃO DESEMPENHO ALGORITMO		**Informações Gerais**	1	Número total de pacientes
			2	Técnica de radioterapia
			3	Método Diagnóstico
			4	Sítio Tumoral
			5	Escopo (segmentação e/ou dosimetria)
			6	Algoritmo de IA utilizado
	MI-CLAIM	**Avaliação de Performance Segmentação**	7	DSC (comparação de desempenho entre a linha de base e o modelo proposto apresentada com a significância estatística apropriada)
			8	HD (comparação de desempenho entre a linha de base e o modelo proposto apresentada com a significância estatística apropriada)
			9	AHD (comparação de desempenho entre a linha de base e o modelo proposto apresentada com a significância estatística apropriada)
	MI-CLAIM ICRU 83 ICRU 91	**Avaliação de Performance Dosimetria PTV**	10	DVHs (comparação de desempenho entre a linha de base e o modelo proposto apresentada com a significância estatística apropriada)
			11	D50% (comparação de desempenho entre a linha de base e o modelo proposto apresentada com a significância estatística apropriada)
			12	Dmédia (comparação de desempenho entre a linha de base e o modelo proposto apresentada com a significância estatística apropriada)
			13	D98% ou Dprox.min ou $D_{V\text{-}35mm3}$ (comparação de desempenho entre a linha de base e o modelo proposto apresentada com a significância estatística apropriada)
			14	D2% ou Dprox.máx ou D_{V35mm3} (comparação de desempenho entre a linha de base e o modelo proposto apresentada com a significância estatística apropriada)
	MI-CLAIM ICRU 83 ICRU 91	**Avaliação de Performance Dosimetria OAR/PRV**	15	Dmédia (comparação de desempenho entre a linha de base e o modelo proposto apresentada com a significância estatística apropriada)
			16	D50% (comparação de desempenho entre a linha de base e o modelo proposto apresentada com a significância estatística apropriada)
			17	D2% ou Dprox.máx ou D_{V35mm3} (comparação de desempenho entre a linha de base e o modelo proposto apresentada com a significância estatística apropriada)
			18	V_D (comparação de desempenho entre a linha de base e o modelo proposto apresentada com a significância estatística apropriada)
			19	HI (comparação de desempenho entre a linha de base e o modelo proposto apresentada com a significância estatística apropriada)
			20	CI (comparação de desempenho entre a linha de base e o modelo proposto apresentada com a significância estatística apropriada)
			21	GI (comparação de desempenho entre a linha de base e o modelo proposto apresentada com a significância estatística apropriada)

	FONTE	PERSPECTIVA	Nr.	DESCRIÇÃO
AVALIAÇÃO VIABILIDADE DE APLICAÇÃO NA PRÁTICA CLÍNICA	*MINIMAR & CHARM*	***Dados***	22	Número de Centros
			23	Regiões ou Locais
			24	Critério de inclusão e exclusão
	MINIMAR, MI-CLAIM, CLAIM & FUTURE-AI	***Confiabilidade dos Dados***	25	A origem dos dados é descrita e o formato original é detalhado.
			26	Descreve os métodos pelos quais os dados foram anonimizados e como as informações de saúde protegidas foram removidos para atender aos EUA (HIPAA), europeus (GDPR) ou outras leis relevantes.
			27	Foi utilizado um mix de imagem de alta variabilidade incluindo imagens de múltiplos centros, diferentes fornecedores e regiões com recursos limitados como os de países de baixa e média renda?
			28	Como os dados faltantes foram gerenciados? *Declara claramente como os dados ausentes foram tratados e quando substituídos por valores aproximados ou previstos descreve os viéses que os dados inseridos podem gerar*
			29	Como os dados foram separados para fins de treinamento, teste e validação?
	CLAIM	***Treinamento***	30	Detalhes da abordagem de treinamento, incluindo técnica para ampliação e diversidade do dado, hiperparâmetros e número de modelos treinados
			31	Método selecionado para o modelo final
	FUTURE-AI	***Rastreabilidade***	32	Foi informado se será disponibilizado um registro estruturado de todo o pipeline de pré-processamento de imagens e dados relacionados?
			33	Foi informado se serão fornecidas as especificações das entradas/saídas, natureza, pré-requisitos e requisitos dos métodos de pré-processamento e preparação de dados?
			34	O modelo possui alguma ferramenta de rastreabilidade que permita monitorar o funcionamento em tempo real para, por exemplo, sinalizar e registrar erros, desvios e degradação de desempenho?
	FUTURE-AI	***Robustez***	35	Solução para homogeneização da imagem em casos de heterogeneidade
			36	Ferramenta de controle de qualidade para identificação de desvios anormais
			37	Relato de incertezas do modelo além do descriminante classificador para cálculo de uma pontuação de confiança
	MI-CLAIM	***Reprodutibilidade***	38	Nível 1: compartilhamento completo do código
			39	Nível 2: permite que um terceiro avalie o código quanto à precisão/equidade; compartilhar os resultados desta avaliação
			40	Nível 3: liberação de uma máquina virtual (binária) para executar o código com novos dados sem compartilhar seus detalhes
			41	Nível 4: sem compartilhamento
	FUTURE-AI	***Explicabilidade***	42	Foram utilizados múltiplos e complementares métodos de explicabilidade?
			43	Quais foram os testes de avaliação qualitativos ou quantitativos utilizados para determinar a robustez e confiabilidade das explicações
			44	Foi avaliado o impacto clínico dos métodos de explicabilidade realizando um estudo de interação humano-IA, onde o profissional de saúde realiza tarefas clínicas utilizando a ferramenta de IA com e sem explicações?
			45	Foi identificado algum viés resultante introduzido pelos métodos de explicabilidade?

4.2.6. Risco de Viés

Para avaliação do risco de viés, será utilizada uma versão adaptada do PROBAST, desenvolvida por meio de um processo Delphi para estudos de radioterapia baseados em IA.

Diante da variabilidade de métricas e ferramentas relatada ao longo deste trabalho, orienta-se que todos os estudos sejam considerados na análise incluindo aqueles com risco de viés intermediário ou alto. Para reduzir a introdução de vieses, recomenda-se atribuir maior peso aos achados de estudos com maior qualidade técnica, ou seja, metodologicamente mais robustos.

Notas: Maiores informações sobre essa versão adaptada estão disponíveis no item "4.1.5. Ferramenta para avaliação de risco de viés".

Recomenda-se a inclusão de estudos com diferentes níveis de risco de viés, mesmo que isso possa aumentar a heterogeneidade dos resultados. Pois, esta estratégia permite uma avaliação mais realista do estado atual da evidência científica, direcionando futuras pesquisas para questões ainda não totalmente exploradas. Isso é particularmente relevante em campos emergentes, onde as metodologias estão em constante evolução. Além disso, ao reconhecer as limitações dos estudos existentes, torna-se possível aprimorar as metodologias e fortalecer a confiabilidade dos resultados futuros.

4.2.7. Síntese de Dados

Em razão da diversidade de métricas tanto na segmentação tumoral quanto no planejamento dosimétrico, acredita-se que não seja possível realizar uma metanálise mesmo adotando um modelo de efeitos aleatórios.

Na impossibilidade de realizar uma análise quantitativa, é recomendado a elaboração de uma síntese narrativa sistemática. Nesse formato, as informações devem ser apresentadas de forma clara em texto e tabelas, com o objetivo de resumir e explicar as características e os resultados dos estudos incluídos. A síntese narrativa deve explorar as relações e os achados tanto dentro de cada estudo quanto entre os diferentes estudos, seguindo as orientações do **CRD** (*Centre for Reviews and Dissemination*, ou Centro de Revisões e Disseminação) (57).

Para viabilizar comparações, é sugerido a realização de uma análise de subgrupo, considerando os seguintes critérios: localização do tumor, método diagnóstico, técnica de radioterapia, escopo (segmentação e/ou planejamento dosimétrico) e algoritmo de IA.

Nota: Outro aspecto relevante relacionado à heterogeneidade, que pode comprometer a metanálise, é o risco de viés. A inclusão de estudos com diferentes níveis de viés pode amplificar os erros já existentes nos estudos individuais, resultando em interpretações equivocadas e comprometendo a validade das conclusões

4.2.8. VIESES SISTÊMICOS

Para o relato seletivo de desfechos, recomenda-se comparar os desfechos relatados no protocolo com os do estudo publicado, caso o protocolo do estudo esteja disponível. Nos casos em que o protocolo não é disponibilizado, orienta-se comparar os desfechos relatados nas seções de métodos com os resultados publicados.

4.2.9. QUALIDADE DO CONJUNTO DE EVIDÊNCIAS

O item “4.1.6 Ferramenta para avaliação da qualidade do estudo e da síntese de evidência” destaca a falta de uma ferramenta padronizada para avaliar a qualidade de estudos na área de inteligência artificial, assim como para a síntese de evidências. Dada a natureza dos estudos “in silico” e as particularidades dessa área, o RATING surge como uma alternativa viável para a avaliação individual dos estudos. Dessa forma, sugere-se a utilização dessa ferramenta, seguida de uma síntese narrativa sistemática, que possibilitará uma análise detalhada e transparente dos resultados.

5. DISCUSSÃO

Na elaboração deste protocolo, emergiram ***dois temas críticos***, que ***antecedem a questão da IA*** e que tornam a avaliação da performance dos algoritmos desafiadora. Ambos, já eram discutidos pela comunidade científica há algum tempo, são eles: a ausência de padronização na prescrição e relato de dose entre as múltiplas diretrizes clínicas (12),(41),(58),(59),(60) e as múltiplas diretrizes de contorno(61),(62),(45),

(46). Com o avanço da IA, a relevância desses aspectos se intensifica, dado o impacto direto que ambos exercem tanto na generalização quanto na validação dos modelos. Na **Figura 11**, são destacados os impactos tanto na perspectiva clínica quanto no desenvolvimento de algoritmos.

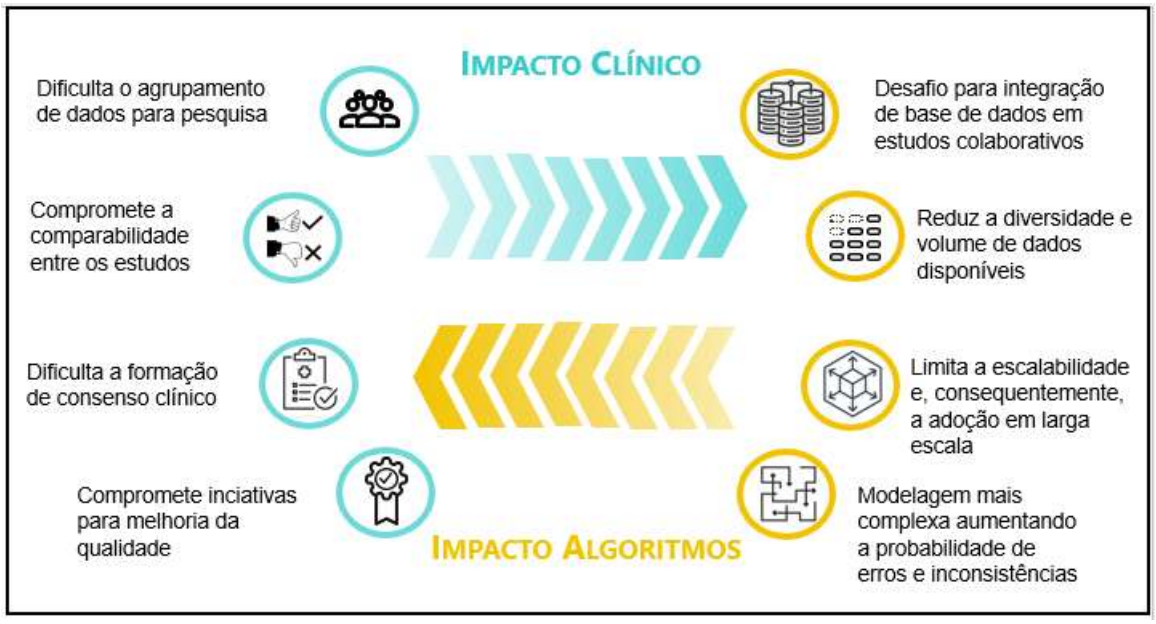


Figura 11: Ausência de padronização – Impactos na perspectiva clínica e no desenvolvimento de algoritmos

Ilustração criada pelo autor com base nas seguintes referências bibliográficas (12), (41), (58), (59) e (60)

Quando a estruturação do dado é inconsistente e limitada, aumenta o custo, a complexidade e o esforço para extrair informação de valor clínico (63). A padronização oferece incentivos para a agregação de informações e apoia a criação de soluções comuns entre fornecedores, promovendo a interoperabilidade e a integração de dados multicêntricos. Além disso, do ponto de vista clínico, a adoção de diretrizes aumenta a exatidão do tratamento, melhora os desfechos clínicos e reduz a toxicidade (46).

No que se refere à prescrição e relato de dose, reforça-se a importância da padronização através da adoção das diretrizes ICRU (12) (40) (41). No entanto, a complexidade do tema exige consenso e um escopo bem definido para evitar impasses e a proliferação de métricas comprometendo a eficácia da padronização (58). Uma abordagem gradual, iniciando com um conjunto básico de dados e expandindo-o gradualmente à medida que a comunidade científica enxergue o valor, pode ser uma alternativa. A colaboração entre instituições e sociedades como **AAPM** (American Association of Physicists in Medicine ou Associação Americana de Físicos

em Medicina), AOCR, ASTRO e ESTRO em projetos de agregação de dados é essencial para o desenvolvimento de padrões globais e a criação de soluções que atendam às necessidades da comunidade radioterápica e a adoção de padrões por fornecedores (63).

Quanto à segmentação, existe um duplo desafio para manejo da heterogeneidade dos dados. Um pré-existente, relacionado à falta de consenso em relação às diretrizes de contorno; e um novo, introduzido pela IA, representado pela ausência de métricas padronizadas para avaliação de performance dos algoritmos. Aqui, são abordadas possíveis alternativas para condução de ambos os cenários assim como a intersecção entre eles.

A carência de consenso metodológico, aliada às particularidades de cada instituição, resulta num cenário de múltiplas diretrizes para uma mesma doença, tornando a escolha e aplicação mais complexas. Numa revisão sistemática realizada em 2020, foram identificadas 142 diretrizes de contorno. Essa revisão teve como foco diretrizes de contorno para radioterapia, utilizando dados provenientes da rede internacional de diretrizes. A região de cabeça e pescoço destacou-se como a área com o maior número de diretrizes, totalizando 34, conforme ilustrado na **Figura 12**

Figura 12: Distribuição das diretrizes de contorno conforme localização do tumor (46)(46).

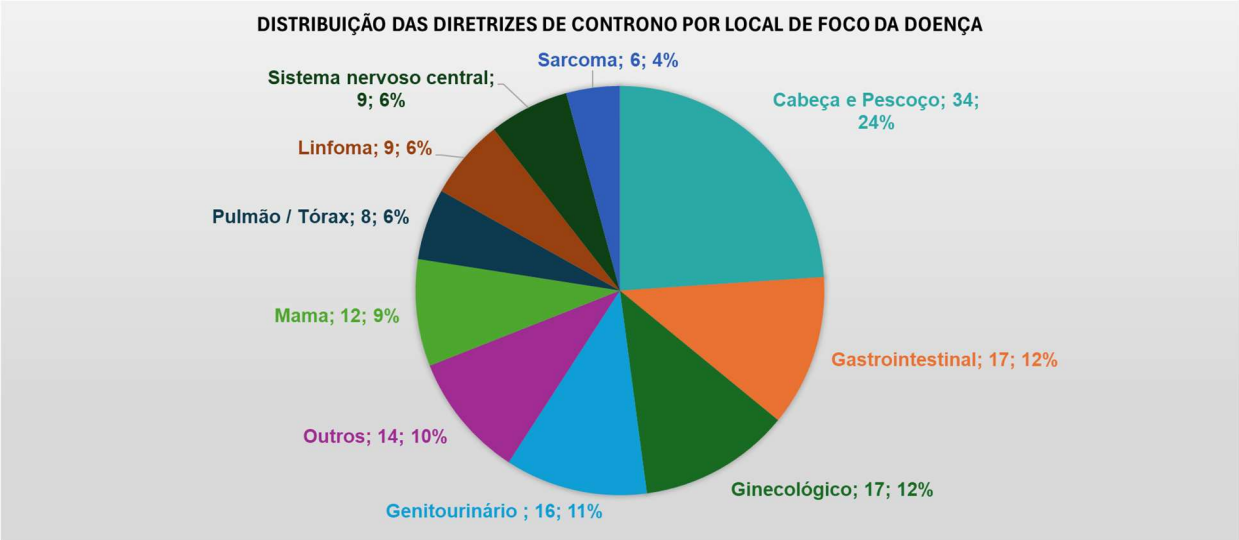


Figura 12: Distribuição das diretrizes de contorno conforme localização do tumor (46)

A despeito da crescente importância das imagens 3D na radioterapia, apenas 16% das diretrizes identificadas incluíam um conjunto completo de imagens de casos e a maioria dessas estavam no contexto de garantia da qualidade em ensaios clínicos.

A soma desses fatores dificulta a disseminação e a utilização das diretrizes. Os clínicos demonstram preferência por recursos de contorno baseados em imagens, mais acessíveis e fáceis de usar, para melhorar a prática clínica (46).

Em virtude do exposto, para superar os desafios relativos à segmentação e em linha com a preferência expressa pelos clínicos no parágrafo acima, a diretriz conjunta da ESTRO e AAPM para o desenvolvimento, validação clínica e relato de modelos de inteligência artificial em radioterapia recomenda investir no desenvolvimento e disseminação pública, ou acesso cego, a conjuntos de dados de referência (7).

Quanto ao segundo desafio que surge com a IA, definição de métricas de performance, essa mesma diretriz recomenda o relato dos parâmetros DSC e HD. Dada a natureza puramente matemática dessas métricas, certamente, não serão suficientes para prover a segurança necessária para adoção clínica, visto que não levam em conta as consequências clínicas das eventuais discrepâncias geométricas (7). Ou seja, o tratamento insuficiente no caso de uma subavaliação; ou de danos a tecidos saudáveis na ocorrência de uma avaliação superestimada (64).Sendo assim, um conjunto de dados de referência se torna ainda mais imprescindível para o estabelecimento de um padrão de comparação viabilizando a atenuação da

variabilidade intra e interinstitucional existente. Desse modo, estabelece-se um ambiente mais uniforme viabilizando o teste e a validação sistemáticos de novos algoritmos com dados padronizados e consistentes entre os estudos.

Este estudo recomenda a utilização da métrica AVD, ao invés do HD. Tal indicação visa mitigar a sensibilidade a “outliers” (pequenas segmentações adicionais de objetos fora do objeto principal). Maiores detalhes sobre o critério de decisão estão disponíveis no item “4.1.2 Seleção das métricas para avaliação de segmentação”.

Caso o compartilhamento de dados de imagem anonimizados entre múltiplos centros seja inviável, existe a possibilidade de se recorrer a recursos como o aprendizado federado (AF), que surge como uma alternativa promissora para os desenvolvedores de IA em parceria com esses centros. Essa abordagem possibilita o treinamento descentralizado de redes neurais profundas entre vários clientes, preservando a privacidade dos dados de cada instituição conforme ilustrado na **Figura 13** (65).

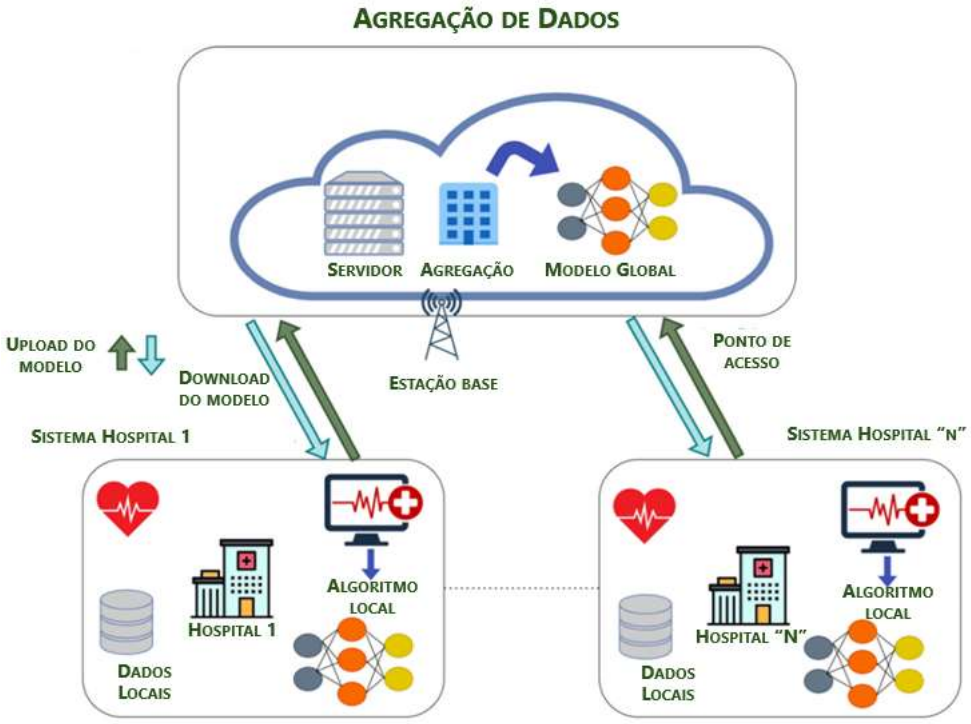


Figura 13: Aprendizado Federado (AF) na área da saúde (66)

Entretanto, a implementação em larga escala do AF ainda enfrenta desafios significativos. A heterogeneidade dos dados pode resultar em soluções globais que não atendam de forma ideal às necessidades de algumas instituições, tornando

fundamental o consenso sobre critérios de otimização antes do treinamento (67). Além disso, tanto o alto custo computacional da transmissão de grandes volumes de dados quanto a restrição de poder computacional nos clientes representam desafios para a escalabilidade do aprendizado federado. A primeira limita a comunicação entre os clientes, enquanto a segunda exige o desenvolvimento de modelos mais eficientes e adaptáveis. Garantir a segurança e a privacidade dos dados também é indispensável, exigindo protocolos criptográficos robustos para proteger as informações durante o processo de treinamento (68).

Apesar desses desafios, o aprendizado federado é uma área de pesquisa dinâmica e promissora, com potencial para transformar a medicina. Sua capacidade de criar modelos mais precisos e personalizados pode impulsionar a medicina de precisão e contribuir para a melhoria dos resultados clínicos (67).

Uma iniciativa relevante no campo dos estudos multicêntricos e da padronização no planejamento radioterápico é o estudo **ARCHERY** (69) que será realizado em países de baixa e média renda. Trata-se de uma pesquisa multicêntrica prospectiva não randomizada realizada em seis hospitais públicos especializados em oncologia na Índia (n=2), Jordânia (n=1), Malásia (n=1) e África do Sul (n=2). Com duração de quatro anos e protocolo publicado em outubro de 2023, o estudo busca avaliar a aceitabilidade geral de planos de tratamento radioterápicos automatizados para cânceres de colo do útero, cabeça e pescoço, e próstata. Isso inclui a análise de volumes-alvo clínicos (CTVs) e órgãos de risco (OARs), com base em diretrizes internacionais de contorno e restrições dosimétricas. O protocolo não informa quais métricas serão utilizadas para avaliação da performance.

A avaliação será coordenada pelo **RTTQA** (*Radiotherapy Trials Quality Assurance* ou Garantia de Qualidade dos Ensaios de Radioterapia) e seguirá padrões internacionais estabelecidos pelo **GHG** (*Global Quality Assurance of Radiation Therapy Clinical Trials Harmonisation Group* ou Grupo de Harmonização Global de Garantia de Qualidade em Ensaios Clínicos de Radioterapia). Um painel internacional de oncologistas radioterápicos com mais de 10 anos de experiência revisará os contornos e planos, utilizando protocolos específicos para volumes-alvo e dosimetria. O estudo

utilizará o **RPA** (*Radiotherapy Planning Assistant* ou Assistente de Planejamento de Radioterapia), um software de inteligência artificial desenvolvido pelo MD Anderson Cancer Center, projetado para automatizar tarefas de contorno e realizar verificações de garantia de qualidade. O cálculo de dose e a otimização dos planos serão realizados pelo sistema Eclipse (Varian Medical Systems).

Se bem-sucedido, o software será disponibilizado como um serviço web sem fins lucrativos para hospitais públicos em países de baixa e média renda, promovendo o acesso a tratamentos radioterápicos de alta qualidade e ampliando a capacidade de atendimento ao câncer nessas regiões.

Outro aspecto extremamente relevante que surgiu durante a elaboração do protocolo foi a *dificuldade em definir as ferramentas* mais apropriadas para *avaliação da aplicabilidade clínica e risco de viés.*

Começando pelas ferramentas para avaliação de aplicabilidade, o alto grau de heterogeneidade das imagens médicas exige um rigoroso processo de avaliação para a incorporação de modelos de IA na prática clínica. Já existem algumas diretrizes específicas de relato nessa área, como CLAIM, **CLEAR** (*CheckList for EvaluAtion of Radiomics Research* ou Lista de Verificação para Avaliação de Pesquisas em Radiômica) (70) e diretrizes voltadas para desenvolvedores, como FUTURE-AI. No entanto, a falta de padronização entre essas diretrizes ressalta a necessidade de um esforço colaborativo para estabelecer uma diretriz única que contemple não apenas a performance dos modelos, mas também sua capacidade de generalização e sua aplicabilidade em diferentes contextos clínicos.

A definição de um conjunto comum de critérios é essencial para garantir tanto a confiabilidade quanto a reprodutibilidade dos resultados, favorecendo a integração dessas tecnologias na prática clínica. A sugestão de aplicar diretrizes combinadas, quando pertinente para indicações sobrepostas (6), introduz vulnerabilidades no processo de avaliação, especialmente se os critérios utilizados não forem respaldados por um consenso entre especialistas e se não forem empregadas medidas estatísticas objetivas para avaliar o impacto da subjetividade dos avaliadores nos resultados. Ainda que o consenso entre especialistas e a análise estatística sejam realizados em cada estudo, o que na prática pode não ser totalmente viável devido a fatores como

recursos e tempo, a comparabilidade entre diferentes pesquisas continuará comprometida, uma vez que a falta de homogeneidade nos critérios de avaliação empregados inviabiliza a padronização dos resultados. Por isso a relevância de padronizar ao máximo os critérios de avaliação.

Em relação ao risco de viés, como já antecipado, o PROBAST é voltado especificamente para modelos de predição em contextos de prognósticos e diagnósticos. Para evitar incompatibilidades com certos termos específicos desses dois casos de uso, foi utilizada uma versão adaptada do PROBAST (50). A confiabilidade da concordância entre os avaliadores foi avaliada por meio da métrica estatística Fleiss' Kappa, cujo resultado foi baixo. Ainda assim, considerando o conjunto de ferramentas disponíveis, essa adaptação se mostrou a opção mais adequada.

Está em curso uma iniciativa destinada a adaptar o PROBAST para torná-lo aplicável a estudos de IA, **PROBAST+AI** (71). A iniciativa é acompanhada com interesse pelo potencial de contribuição para o avanço na avaliação e redução de vieses.

O **PRISMA-AI** (*Preferred Reporting Items for Systematic Reviews and Meta-Analyses tailored for Artificial Intelligence* ou Itens Preferenciais de Relato para Revisões Sistemáticas e Meta-análises adaptados para Inteligência Artificial) é outra ferramenta que desperta grande expectativa. Surge como uma resposta à crescente demanda por diretrizes de relato mais específicas para estudos com inteligência artificial. A explosão de interesse em torno da IA tem evidenciado a necessidade de padrões mais rigorosos para garantir a qualidade e a reprodutibilidade dessas pesquisas. Questões como a explicabilidade dos algoritmos, a eficácia clínica e a padronização no relato dos estudos primários têm se destacado como desafios a serem superados (72).

A explicabilidade tem se consolidado como um tema cada vez mais recorrente em artigos de inteligência artificial voltados para o setor de saúde (13) (21) (26) (73) (74), e essa tendência não apresenta sinais de desaceleração. Um dos elementos que reforça essa tendência, é o fato de agências reguladoras estarem publicando diretrizes que reforçam a necessidade e importância do tema.

Em junho de 2024, por exemplo, o **FDA** (*Food and Drug Administration* ou Administração de Alimentos e Medicamentos), o Health Canada e a **MHRA** (*Medicines*

and Healthcare Products Regulatory Agency ou Agência Reguladora de Medicamentos e Produtos de Saúde), agências reguladoras dos Estados Unidos, Canadá e Reino Unido, respectivamente, publicaram em conjunto o Transparency for Machine Learning-Enabled Medical Devices: Guiding Principles (75) (Transparência para Dispositivos Médicos Habilitados por Aprendizado de Máquina: Princípios Orientadores). Esse documento, que complementa o **GMLP** (76) (*Good Machine Learning Practice for Medical Device Development: Guiding Principles* ou Boas Práticas de Aprendizado de Máquina para o Desenvolvimento de Dispositivos Médicos: Princípios Orientadores) publicado em 2021, foca em princípios orientadores para promover a transparência.

O referido documento descreve a necessidade de comunicar de forma transparente as informações relevantes aos públicos envolvidos. Nesse contexto, a lógica do modelo - ou seja, o raciocínio que embasa as decisões - ganha destaque. A explicabilidade se configura como o grau em que essa lógica pode ser compreendida por um indivíduo, permitindo que ele acompanhe o processo de tomada de decisão do modelo.

O conceito de explicabilidade foi abordado no item “2.3 Os desafios para incorporação na prática clínica” e representa um pilar fundamental para a adoção de sistemas de IA na saúde. Trata-se de um campo de pesquisa amplo e em constante evolução. Neste estudo, o objetivo não é esgotar o tema, mas é fundamental compreender alguns desafios que ainda precisam ser superados para que essa área possa gerar contribuições cada vez mais relevantes. Alguns deles são:

> *Conceito e construção de explicação ideal*: definir o que constitui uma explicação ideal é uma questão complexa em XAI, posto que ainda não há um consenso sobre os critérios que determinam o que seria essa qualidade ideal. Além disso, a criação de explicações de referência a partir da perspectiva humana é inviável, pois isso pressupõe uma compreensão completa dos modelos de DL que, por sua própria natureza, não são transparentes. Uma alternativa viável, seria a criação de conjuntos de dados sintéticos com explicações previamente conhecidas. Contudo, surge o desafio de garantir a representatividade desses dados. É preciso assegurar que os problemas

sintéticos capturem a complexidade e as nuances dos problemas reais para que as conclusões obtidas a partir deles sejam generalizáveis (77).

Métricas: A ausência de métricas adequadas compromete a conexão de um modelo a um conceito concreto de explicabilidade. Para estabelecer uma base sólida de comparação entre diferentes modelos e garantir que as explicações geradas atendam aos requisitos de XAI, é fundamental o desenvolvimento de métricas adequadas. As métricas atuais consideram aspectos como a qualidade da explicação, sua utilidade para o usuário, seu impacto na compreensão do modelo e na confiança do usuário. No entanto, a literatura científica, ainda, aponta para a necessidade de métricas mais quantificáveis e generalizáveis para avaliar de forma precisa e consistente os algoritmos de explicabilidade (16).

Personalização da explicação: criar explicações ajustadas ao usuário diferenciando níveis de conhecimento técnico ou preferências individuas é essencial para facilitar o entendimento. Essa abordagem, baseada em conhecimentos prévios e necessidades específicas, requer uma integração de diversas áreas, como: *engenharia do conhecimento* focada no desenvolvimento de estrutura e sistemas que organizam o conhecimento de modo lógico e compreensível; *ciências cognitivas* que estuda como as pessoas processam informações e tomam decisões; *interface homem-computador* que desenvolve formas acessíveis de apresentar explicações utilizando elementos como visualizações, interatividade e design amigável. Essa sinergia entre diferentes disciplinas resulta numa experiência de aprendizado mais eficaz que aprimora a qualidade da tomada de decisão e democratiza o acesso à informação (77).

Este trabalho apresenta algumas limitações. A primeira está relacionada aos critérios para avaliação da aplicação na prática clínica os quais foram desenvolvidos de maneira não sistemática, através da combinação de múltiplas ferramentas sem adoção de um protocolo estruturado. A ausência de um painel de especialistas para validar os itens do questionário e a falta de aplicação de testes estatísticos para avaliar a subjetividade dos avaliadores representam limitações metodológicas que podem influenciar resultados e limitar a generalização das conclusões. Esses fatores

comprometem a comparabilidade dos resultados com outros estudos. A segunda refere-se à adaptação da ferramenta do PROBAST. O grupo que realizou a referida adequação, utilizou uma métrica estatística, Fleiss' kappas, para avaliar a confiabilidade da concordância e o resultado foi baixo (50). Apesar disso, sem a adequação de terminologias propostas neste trabalho as quais foram indicadas na seção "4.1.5 Ferramenta para avaliação de risco de viés" haveria uma incompatibilidade entre os termos utilizados no contexto de prognóstico individual ou predição diagnóstica e aqueles empregados na avaliação do desempenho da IA em oncologia por radiação. Logo, dentro do arcabouço de ferramentas disponíveis, essa foi a melhor opção encontrada. Por fim, a recomendação para inclusão de estudos primários com risco de viés moderado ou alto compromete a robustez e a credibilidade da síntese dos resultados. Contudo, dado o estágio emergente da tecnologia, é fundamental identificar limitações dos estudos existentes para aprimorar as metodologias e fortalecer a robustez dos resultados em futuras pesquisas.

6. **CONCLUSÃO**

Para que a utilização da IA na radioterapia possa reduzir as disparidades entre países de alta e média/baixa renda, é crucial universalizar o acesso e a compreensão destes sistemas. Isso exige algoritmos baseados em padrões robustos, o que demanda um esforço conjunto de diversos atores. Médicos e físicos médicos, por meio de suas associações, devem definir diretrizes clínicas; hospitais precisam compartilhar dados de imagem e criar bancos de dados públicos; desenvolvedores necessitam dedicar-se a projetar e documentar mecanismos que garantam a segurança, a transparência e a reprodutibilidade das ferramentas; e editores e revisores devem assegurar a qualidade dos relatos e o alinhamento com padrões emergentes. Estes últimos, em conjunto, com pesquisadores, metodologistas e organizações internacionais podem propor o desenvolvimento de diretrizes e ferramentas a partir da identificação de necessidades específicas. Embora avanços significativos já tenham sido alcançados, o uso da IA na radioterapia ainda é incipiente. A padronização de métricas e ferramentas é fundamental para validar o desempenho dos sistemas e acelerar o desenvolvimento tecnológico. Sem isso, os esforços no desenvolvimento de

ferramentas e propostas metodológicas de avaliação estarão sempre sujeitos a inconsistências que comprometem a sua aplicação e validação.

7. REFERÊNCIAS BIBLIOGRÁFICAS

1. Arthur Accioly Rosa, Homero Lavieri Martins, Leonardo Pimentel, Marcus Simões Castilho, Silvério Marinho, Virginia Izabel Oliveira. Projeto RT2030: Planejamento de Desenvolvimento da Radioterapia para a Próxima Década [Internet]. Sociedade Brasileira de Radioterapia. 2021 [cited 2021 Nov 10]. Available from: https://sbradioterapia.com.br/wp-content/uploads/2021/08/Relatorio_Projeto_RT2030.pdf

2. Rosenblatt E, Zubizarreta E. RADIOTHERAPY IN CANCER CARE: FACING THE GLOBAL CHALLENGE.

3. Sheng K. Artificial intelligence in radiotherapy: a technological review. Front Med. 2020;14(4):431–49.

4. Chen Z, King W, Pearcey R, Kerba M, Mackillop WJ. The relationship between waiting time for radiotherapy and clinical outcomes: a systematic review of the literature. Radiother Oncol [Internet]. 2008 Apr [cited 2021 Nov 10];87(1):3–16. Available from: https://pubmed.ncbi.nlm.nih.gov/18160158/

5. Direcção-Geral da Saúde. Estrutura Concetual da Classificação Internacional sobre Segurança do Doente. Relatório técnico Final [Internet]. 2011;142. Available from: http://www.dgs.pt/documentos-e-publicacoes/classificacao-internacional-sobre-seguranca-do-doente.aspx

6. Klontzas ME, Gatti AA, Tejani AS, Kahn CE. AI Reporting Guidelines: How to Select the Best One for Your Research. Radiol Artif Intell. 2023;5(3).

7. Hurkmans C, Bibault JE, Brock KK, van Elmpt W, Feng M, David Fuller C, et al. A joint ESTRO and AAPM guideline for development, clinical validation and reporting of artificial intelligence models in radiation therapy. Radiother Oncol [Internet]. 2024;197(May):110345. Available from: https://doi.org/10.1016/j.radonc.2024.110345

8. Sarkar R, Samuel D, Dunbar L, Monnerat G. 5 years of The Lancet Digital Health. Lancet Digit Heal [Internet]. 2024;6(5):e299. Available from: http://dx.doi.org/10.1016/S2589-7500(24)00073-6

9. Kolbinger FR, Veldhuizen GP, Zhu J, Truhn D, Kather JN. Reporting guidelines in medical artificial intelligence: a systematic review and meta-analysis. Commun Med. 2024;4(1):1–10.

10. Moons K, Kaul T, Kaul T. PROBAST+AI Delphi Survey Plan Overview. 2024;(April):1–10.

11. Taha AA, Hanbury A. Metrics for evaluating 3D medical image segmentation: Analysis, selection, and tool. BMC Med Imaging [Internet]. 2015;15(1). Available from: http://dx.doi.org/10.1186/s12880-015-0068-x

12. Das IJ, Yadav P, Andersen AD, Chen ZJ, Huang L, Langer MP, et al. Dose prescription and reporting in stereotactic body radiotherapy: A multi-institutional study. Radiother Oncol. 2023;182:1–20.

13. Kapoor N, Lacson R, Khorasani R. Workflow Applications of Artificial Intelligence in Radiology and an Overview of Available Tools. J Am Coll Radiol [Internet]. 2020 [cited 2022 Nov 30];17:1363–70. Available from: https://doi.org/10.1016/j.jacr.2020.08.016

14. Lekadir K, Osuala R, Gallin C, Lazrak N, Kushibar K, Tsakou G, et al. FUTURE-AI : Guiding Principles and Consensus Recommendations for Trustworthy Artificial Intelligence in. 2021;(ii).

15. Topol EJ. Welcoming new guidelines for AI clinical research. Nat Med. 2020;26(9):1318–20.

16. Barredo A, Díaz-rodríguez N, Del J, Bennetot A, Tabik S, Barbado A, et al. Explainable Artificial Intelligence (XAI): Concepts , taxonomies , opportunities and challenges toward responsible AI. Inf Fusion [Internet]. 2020;58(December 2019):82–115. Available from: https://doi.org/10.1016/j.inffus.2019.12.012

17. Mukhamediev RI, Symagulov A, Kuchin Y, Yakunin K, Yelis M. From classical machine learning to deep neural networks: A simplified scientometric review. Appl Sci. 2021;11(12).

18. Mukhamediev RI, Popova Y, Kuchin Y, Zaitseva E, Kalimoldayev A, Symagulov A, et al. Review of Artificial Intelligence and Machine Learning Technologies:

Classification, Restrictions, Opportunities and Challenges. Mathematics. 2022;10(15):1–25.

19. Nguyen G, Dlugolinsky S, Bobák M, Tran V, López García Á, Heredia I, et al. Machine Learning and Deep Learning frameworks and libraries for large-scale data mining: a survey. Artif Intell Rev [Internet]. 2019;52(1):77–124. Available from: https://doi.org/10.1007/s10462-018-09679-z

20. Miotto R, Wang F, Wang S, Jiang X, Dudley JT. Deep learning for healthcare: Review, opportunities and challenges. Brief Bioinform. 2017;19(6):1236–46.

21. Huynh E, Hosny A, Guthier C, Bitterman DS, Petit SF, Haas-Kogan DA, et al. Artificial intelligence in radiation oncology. Nat Rev Clin Oncol [Internet]. 2020;17(12):771–81. Available from: http://dx.doi.org/10.1038/s41571-020-0417-8

22. Wang M, Zhang Q, Lam S, Cai J, Yang R. A Review on Application of Deep Learning Algorithms in External Beam Radiotherapy Automated Treatment Planning. Front Oncol. 2020;10(October).

23. Yoriyaz H, Fonseca GP, Bellezzo M. Sistemas de Planejamento em Radioterapia. Rev Bras Física Médica. 2019;13(1):92.

24. Sarrut D, Etxebeste A, Muñoz E, Krah N, Létang JM. Artificial Intelligence for Monte Carlo Simulation in Medical Physics. Front Phys. 2021;9(October):1–13.

25. Bajwa J, Munir U, Nori A, Williams B. Artificial intelligence in healthcare: transforming the practice of medicine. Futur Healthc J. 2021;8(2):e188–94.

26. Galić I, Habijan M, Leventić H, Romić K. Machine Learning Empowering Personalized Medicine: A Comprehensive Review of Medical Image Analysis Methods. Electron. 2023;12(21).

27. Xiang Q, Zi L, Cong X, Wang Y. Concept Drift Adaptation Methods under the Deep Learning Framework: A Literature Review. Appl Sci. 2023;13(11).

28. Lu J, Liu A, Dong F, Gu F, Gama J, Zhang G. Learning under Concept Drift: A Review. IEEE Trans Knowl Data Eng. 2019;31(12):2346–63.

29. Kremers R. Artifi cial Intelligence. Lev Des. 2020;341–68.

30. Guidotti R, Monreale A, Pedreschi D, Giannotti F. Principles of explainable artificial intelligence. Explain AI Within Digit Transform Cyber Phys Syst XAI Methods Appl. 2021;9–31.

31. Saranya A, Subhashini R. A systematic review of Explainable Artificial Intelligence models and applications: Recent developments and future trends. Decis Anal J [Internet]. 2023;7(April):100230. Available from: https://doi.org/10.1016/j.dajour.2023.100230

32. Kamioka H. Preferred reporting items for systematic review and meta-analysis protocols (prisma-p) 2015 statement. Japanese Pharmacol Ther. 2019;47(8):1177–85.

33. Cilla S, Deodato F, Romano C, Ianiro A, Macchia G, Re A, et al. Personalized automation of treatment planning in head-neck cancer: A step forward for quality in radiation therapy? Phys MEDICA-EUROPEAN J Med Phys. 2021;82:7–16.

34. Beavis AW. Treatment Planning Challenges in Breast Irradiation: The Ideal and The Practical. Clin Oncol. 2006;18.

35. Narisada K. Advanced Radiation Treatment Planning of Prostate Cancer. Middle East J Cancer [Internet]. 2018 Apr 3 [cited 2022 Nov 30];6(3):199–200. Available from: https://www.intechopen.com/state.item.id

36. Nai YH, Teo BW, Tan NL, O'Doherty S, Stephenson MC, Thian YL, et al. Comparison of metrics for the evaluation of medical segmentations using prostate MRI dataset. Comput Biol Med [Internet]. 2021;134:104497. Available from: https://doi.org/10.1016/j.compbiomed.2021.104497

37. Müller D, Soto-Rey I, Kramer F. Towards a guideline for evaluation metrics in medical image segmentation. BMC Res Notes. 2022;15(1):1–7.

38. Roberts JS, Montoya LN. Mitigating Bias in Machine Learning. 2024;

39. Hernandez V, Hansen CR, Widesott L, Bäck A, Canters R, Fusella M, et al. What is plan quality in radiotherapy? The importance of evaluating dose metrics, complexity, and robustness of treatment plans. Radiother Oncol [Internet].

2020;153:26–33. Available from: https://doi.org/10.1016/j.radonc.2020.09.038

40. Das IJ, Cheng CW, Chopra KL, Mitra RK, Srivastava SP, Glatstein E. Intensity-modulated radiation therapy dose prescription, recording, and delivery: Patterns of variability among institutions and treatment planning systems. J Natl Cancer Inst. 2008;100(5):300–7.

41. Das IJ, Andersen A, Chen Z (Jay), Dimofte A, Glatstein E, Hoisak J, et al. State of dose prescription and compliance to international standard (ICRU-83) in intensity modulated radiation therapy among academic institutions. Pract Radiat Oncol [Internet]. 2017;7(2):e145–55. Available from: http://dx.doi.org/10.1016/j.prro.2016.11.003

42. Wilke L, Andratschke N, Blanck O, Brunner TB, Combs SE, Grosu AL, et al. ICRU report 91 on prescribing, recording, and reporting of stereotactic treatments with small photon beams: Statement from the DEGRO/DGMP working group stereotactic radiotherapy and radiosurgery. Strahlentherapie und Onkol. 2019;195(3):193–8.

43. Menzel HG. The international commission on radiation units and measurements. J ICRU. 2010;10(1):1–106.

44. International THE, On C, Units R. Prescribing, recording, and reporting of stereotactic treatments with small photon beams. J ICRU. 2014;14(2):1–160.

45. Wright JL, Yom SS, Awan MJ, Dawes S, Fischer-Valuck B, Kudner R, et al. Standardizing Normal Tissue Contouring for Radiation Therapy Treatment Planning: An ASTRO Consensus Paper. Pract Radiat Oncol [Internet]. 2019;9(2):65–72. Available from: https://doi.org/10.1016/j.prro.2018.12.003

46. Kristi L. Stringer, Bulent Turan, Lisa McCormick, Modupeoluwa Durojaiye, Laura Nyblade, Mirjam-Colette Kempf, Bronwen Lichtenstein and JMT. A Systematic Review of Contouring Guidelines in Radiation Oncology: Analysis of Frequency, Methodology, and Delivery of Consensus Recommendations. Physiol Behav. 2017;176(3):139–48.

47. Mongan J, Moy L, Kahn CE. Checklist for Artificial Intelligence and Medical

Imaging (Claim). Radiol Artif Intell. 2020;

48. Hernandez-Boussard T, Bozkurt S, Ioannidis JPA, Shah NH. MINIMAR (MINimum information for medical AI reporting): Developing reporting standards for artificial intelligence in health care. J Am Med Informatics Assoc. 2020;27(12):2011–5.

49. Collins GS, Moons KGM, Dhiman P, Riley RD, Beam AL, Van Calster B, et al. TRIPOD+AI statement: Updated guidance for reporting clinical prediction models that use regression or machine learning methods. Bmj. 2024;

50. Hurkmans C, Bibault JE, Clementel E, Dhont J, van Elmpt W, Kantidakis G, et al. Assessment of bias in scoring of AI-based radiotherapy segmentation and planning studies using modified TRIPOD and PROBAST guidelines as an example. Radiother Oncol [Internet]. 2024;194(February):110196. Available from: https://doi.org/10.1016/j.radonc.2024.110196

51. Hansen CR, Crijns W, Hussein M, Rossi L, Gallego P, Verbakel W, et al. Radiotherapy Treatment plannINg study Guidelines (RATING): A framework for setting up and reporting on scientific treatment planning studies. Radiother Oncol [Internet]. 2020;153:67–78. Available from: https://doi.org/10.1016/j.radonc.2020.09.033

52. Comissão Global de Evidências para Responder aos Desafios Sociais. O relatório da Comissão de Evidências: Um chamado para a ação e caminho a seguir para tomadores de decisão, intermediários de evidências e produtores de evidências orientadas para o impacto [Internet]. 2022. Available from: https://www.mcmasterforum.org/networks/evidence-commission

53. Balshem H, Helfand M, Schünemann HJ, Oxman AD, Kunz R, Brozek J, et al. GRADE guidelines: 3. Rating the quality of evidence. J Clin Epidemiol. 2011;64(4):401–6.

54. AIM Specialty Health. Advanced Imaging Appropriate Use Criteria: Oncologic Imaging, Clinical Appropriateness Guidelines. 2019;100. Available from: www.aimspecialtyhealth.comAppropriate.Safe.Affordable

55. Lenchik L, Heacock L, Weaver AA, Boutin RD, Cook TS, Itri J, et al. Automated Segmentation of Tissues Using CT and MRI: A Systematic Review. :1695–706.

56. Moreira M, Hespanhol R, Leite J. PET/TC em câncer de pulmão: indicações, achados e perspectivas futuras. 2016;25(2):38–40.

57. Ummah MS. CRD's Guidance for Undertaking Reviews in Healthcare [Internet]. Vol. 11, Sustainability (Switzerland). 2019. 1–14 p. Available from: http://scioteca.caf.com/bitstream/handle/123456789/1091/RED2017-Eng-8ene.pdf?sequence=12&isAllowed=y%0Ahttp://dx.doi.org/10.1016/j.regsciurbeco.2008.06.005%0Ahttps://www.researchgate.net/publication/305320484_SISTEM_PEMBETUNGAN_TERPUSAT_STRATEGI_MELESTARI

58. Evans SB, Fraass BA, Berner P, Collins KS, Nurushev T, O'Neill MJ, et al. Standardizing dose prescriptions: An ASTRO white paper. Pract Radiat Oncol [Internet]. 2016;6(6):e369–81. Available from: http://dx.doi.org/10.1016/j.prro.2016.08.007

59. Mayo CS, Moran JM, Bosch W, Xiao Y, McNutt T, Popple R, et al. American Association of Physicists in Medicine Task Group 263: Standardizing Nomenclatures in Radiation Oncology. Int J Radiat Oncol Biol Phys [Internet]. 2018;100(4):1057–66. Available from: https://doi.org/10.1016/j.ijrobp.2017.12.013

60. Mohan A, Forde E. Adherence to ICRU-83 reporting recommendations is inadequate in prostate dosimetry studies. Pract Radiat Oncol [Internet]. 2018;8(3):e133–8. Available from: http://dx.doi.org/10.1016/j.prro.2017.08.006

61. Grégoire V, Ang K, Budach W, Grau C, Hamoir M, Langendijk JA, et al. Delineation of the neck node levels for head and neck tumors: A 2013 update. DAHANCA, EORTC, HKNPCSG, NCIC CTG, NCRI, RTOG, TROG consensus guidelines. Radiother Oncol. 2014;110(1):172–81.

62. Salembier C, Villeirs G, De Bari B, Hoskin P, Pieters BR, Van Vulpen M, et al. ESTRO ACROP consensus guideline on CT- and MRI-based target volume delineation for primary radiation therapy of localized prostate cancer. Radiother Oncol [Internet]. 2018;127(1):49–61. Available from:

https://doi.org/10.1016/j.radonc.2018.01.014

63. Mayo CS, Kessler ML, Eisbruch A, Weyburne G, Feng M, Hayman JA, et al. The big data effort in radiation oncology: Data mining or data farming? Adv Radiat Oncol [Internet]. 2016;1(4):260–71. Available from: http://dx.doi.org/10.1016/j.adro.2016.10.001

64. Kim H, Monroe JI, Lo S, Yao M, Harari PM, Machtay M, et al. Quantitative evaluation of image segmentation incorporating medical consideration functions. Med Phys. 2015;42(6):3013–23.

65. Darzidehkalani E, Ghasemi-rad M, van Ooijen PMA. Federated Learning in Medical Imaging: Part II: Methods, Challenges, and Considerations. J Am Coll Radiol [Internet]. 2022;19(8):975–82. Available from: https://doi.org/10.1016/j.jacr.2022.03.016

66. Rahman A, Hossain MS, Muhammad G, Kundu D, Debnath T, Rahman M, et al. Federated learning-based AI approaches in smart healthcare: concepts, taxonomies, challenges and open issues [Internet]. Vol. 26, Cluster Computing. Springer US; 2023. 2271–2311 p. Available from: https://doi.org/10.1007/s10586-022-03658-4

67. Rieke N, Hancox J, Li W, Milletarì F, Roth HR, Albarqouni S, et al. The future of digital health with federated learning. npj Digit Med [Internet]. 2020;3(1):1–7. Available from: http://dx.doi.org/10.1038/s41746-020-00323-1

68. Hernandez-Cruz N, Saha P, Sarker MMK, Noble JA. Review of Federated Learning and Machine Learning-Based Methods for Medical Image Analysis. Big Data Cogn Comput. 2024;8(9):99.

69. Aggarwal A, Court LE, Hoskin P, Jacques I, Kroiss M, Laskar S, et al. ARCHERY: a prospective observational study of artificial intelligence-based radiotherapy treatment planning for cervical, head and neck and prostate cancer - study protocol. BMJ Open. 2023;13(12):1–7.

70. Kocak B, Baessler B, Bakas S, Cuocolo R, Fedorov A, Maier-Hein L, et al. CheckList for EvaluAtion of Radiomics research (CLEAR): a step-by-step

reporting guideline for authors and reviewers endorsed by ESR and EuSoMII. Insights Imaging [Internet]. 2023;14(1). Available from: https://doi.org/10.1186/s13244-023-01415-8

71. Collins GS, Dhiman P, Andaur Navarro CL, Ma J, Hooft L, Reitsma JB, et al. Protocol for development of a reporting guideline (TRIPOD-AI) and risk of bias tool (PROBAST-AI) for diagnostic and prognostic prediction model studies based on artificial intelligence. BMJ Open. 2021;11(7):1–7.

72. Cacciamani GE, Chu TN, Sanford DI, Abreu A, Duddalwar V, Oberai A, et al. PRISMA AI reporting guidelines for systematic reviews and meta-analyses on AI in healthcare. Nat Med. 2023;29(1):14–5.

73. Liu X, Faes L, Kale AU, Wagner SK, Fu DJ, Bruynseels A, et al. A comparison of deep learning performance against health-care professionals in detecting diseases from medical imaging: a systematic review and meta-analysis. Lancet Digit Heal. 2019 Oct 1;1(6):e271–97.

74. Magrabi F, Ammenwerth E, Mcnair JB, De Keizer NF, Hyppönen H, Nykänen P, et al. Artificial Intelligence in Clinical Decision Support: Challenges for Evaluating AI and Practical Implications A Position Paper from the IMIA Technology Assessment & Quality Development in Health Informatics Working Group and the EFMI Working Group for Assessment of Health Information Systems. 2019 [cited 2021 Nov 10]; Available from: http://dx.doi.org/10.1055/s-0039-1677903

75. U.S. Food & Drug Administration, Health Canada, Medicines & Healthcare products Regulatory Agency. Transparency for Machine Learning-Enabled Medical Devices: Guiding Principles. 2024;(June).

76. US FDA. Good Machine Learning Practice for Medical Device Development: Guiding Principles. US Food Drug Adm [Internet]. 2021;(October):1. Available from: https://www.fda.gov/medical-devices/software-medical-device-samd/good-machine-learning-practice-medical-device-development-guiding-principles

77. Samek W, Montavon G, Lapuschkin S, Anders CJ, Müller KR. Explaining Deep Neural Networks and Beyond: A Review of Methods and Applications. Proc

IEEE. 2021;109(3):247–78.

INSTITUTO DE PESQUISAS ENERGÉTICAS E NUCLEARES
Diretoria de Pesquisa, Desenvolvimento e Ensino
Av. Prof. Lineu Prestes, 2242 – Cidade Universitária CEP: 05508-000
Fone (11) 2810-1570 ou (11) 2810-1572
SÃO PAULO – São Paulo – Brasil
http://mprofissional.ipen.br

O Instituto de Pesquisas Energéticas e Nucleares (IPEN) é uma Autarquia vinculada à Secretaria de Desenvolvimento Econômico do Governo do Estado de São Paulo e gerida técnica e administrativamente pela Comissão Nacional de Energia Nuclear (CNEN), órgão do Ministério da Ciência, Tecnologia e Inovações (MCTI) do Governo Federal.