INTELIGÊNCIA ARTIFICIAL NO DIAGNÓSTICO POR IMAGEM: APLICAÇÕES, LIMITAÇÕES E PERSPECTIVAS PARA A PRÁTICA MÉDICA

ARTIFICIAL INTELLIGENCE IN DIAGNOSTIC IMAGING: APPLICATIONS, LIMITATIONS, AND PERSPECTIVES FOR MEDICAL PRACTICE

https://doi.org/10.63330/aurumpub.068-074

Élida Lúcia Ferreira Assunção
Odontologia; mestrado em Clínicas Odontológicas — Universidade Estadual de Montes Claros (Unimontes), Montes Claros–MG.
E-mail: draelidalucia@hotmail.com

Juliano dos Santos
Enfermagem — Instituto Nacional de Câncer (INCA), Rio de Janeiro–RJ.
E-mail: juliano.santos@inca.gov.br

Aline Pelosini Gomes
Medicina — Universidade Metropolitana de Santos (UNIMES), Santos–SP.
E-mail: alinegomesapg@gmail.com

Emily Vitória Maia e Souza Beserra
Medicina — Faculdade de Petrolina (FACAPE), Petrolina–PE.
E-mail: emily.beserra25278@aluno.facape.br

Gabrielle da Silva Rosa
Medicina — Fundação Universidade Federal de Rondônia (UNIR), Porto Velho–RO.
E-mail: gabrielesilva.11@gmail.com

Cláudio do Prado Aranha Filho
Medicina — Faculdade Morgana Potrich (FAMP), Goiânia–GO.
E-mail: claudioaranhafilho@gmail.com

Resumo

A inteligência artificial atua em diferentes etapas do diagnóstico por imagem, incluindo aquisição, reconstrução, detecção de alterações, segmentação, quantificação e elaboração de laudos. Este capítulo analisa seus fundamentos e aplicações clínicas, distinguindo desempenho técnico de benefício assistencial. A discussão baseia-se em estudos primários, ensaios clínicos e diretrizes de relato e governança, com atualização até outubro de 2026. São examinados exemplos em mamografia, radiografia torácica, tomografia computadorizada, ressonância magnética e ultrassonografia, além de aplicações em radiômica e processamento de linguagem. A evidência recente demonstra benefícios em tarefas delimitadas, mas também expõe problemas de generalização, viés entre subgrupos, instabilidade da reconstrução e dependência do fluxo de atendimento. Resultados favoráveis de rastreamento mamográfico coexistem com

ausência de redução do tempo diagnóstico em um ensaio de priorização de radiografias, ressaltando a necessidade de avaliar o sistema integrado ao serviço. A implementação exige validação externa, avaliação prospectiva, definição de responsabilidades e monitoramento após a adoção. Conclui-se que o valor clínico da inteligência artificial depende da tarefa, da população, da qualidade dos dados e da interação com os profissionais, sem equivalência entre automação de uma etapa e autonomia diagnóstica irrestrita.

Palavras-chave: Acurácia diagnóstica; Aprendizado profundo; Diagnóstico por imagem; Inteligência artificial; Segurança do paciente.

Abstract

Artificial intelligence plays a role at various stages of diagnostic imaging, including acquisition, reconstruction, detection of abnormalities, segmentation, quantification, and report generation. This chapter examines its foundations and clinical applications, distinguishing between technical performance and clinical benefit. The discussion is based on primary studies, clinical trials, and guidelines for reporting and governance, with information updated as of October 2026. Examples are explored across mammography, chest radiography, computed tomography, magnetic resonance imaging, and ultrasound, alongside applications in radiomics and natural language processing. Recent evidence demonstrates benefits in specific, well-defined tasks, while also highlighting challenges such as poor generalizability, bias across subgroups, reconstruction instability, and dependency on clinical workflow. Favorable results in mammography screening contrast with findings from a trial on radiography prioritization—where no reduction in diagnostic turnaround time was observed—underscoring the need to evaluate these systems within the context of the clinical service. Implementation requires external validation, prospective evaluation, clear definition of responsibilities, and post-adoption monitoring. It is concluded that the clinical value of artificial intelligence depends on the specific task, the patient population, data quality, and the interaction with healthcare professionals; the automation of a single step should not be equated with unrestricted diagnostic autonomy.

Keywords: Artificial intelligence; Deep learning; Diagnostic accuracy; Diagnostic imaging; Patient safety.

1 INTRODUÇÃO

O diagnóstico por imagem combina aquisição de sinais, reconstrução, reconhecimento de padrões e interpretação clínica. A inteligência artificial (IA) pode intervir em cada uma dessas etapas, mas o significado de seu resultado varia: um contorno anatômico, uma probabilidade de malignidade, um alerta

de prioridade e um texto de laudo não respondem à mesma pergunta. Avaliar essas ferramentas requer identificar a tarefa e a decisão que será modificada, antes de comparar medidas de desempenho.

A expansão do aprendizado profundo ampliou a capacidade de representar padrões diretamente a partir de imagens. A U-Net consolidou uma arquitetura para segmentação biomédica, enquanto sistemas como o nnU-Net demonstraram a importância da configuração do processamento e do treinamento, além da arquitetura isolada (Ronneberger; Fischer; Brox, 2015; Isensee et al., 2021). Esses avanços técnicos não eliminam a necessidade de validação clínica, pois o desempenho obtido em uma base de pesquisa pode não se reproduzir em outro serviço.

O problema central é a passagem da capacidade computacional para o benefício assistencial. No rastreamento mamográfico, ensaios prospectivos já avaliaram a organização da leitura com IA; em outras tarefas, a evidência permanece predominantemente retrospectiva ou limitada a testes controlados. Este capítulo discute aplicações relevantes para a prática médica, seus limites metodológicos e os requisitos para adoção, com base em estudos primários e documentos de referência consultados até 26 de outubro de 2026. Trata-se de síntese narrativa, sem pretensão de busca sistemática exaustiva ou estimativa quantitativa combinada.

2 FUNDAMENTOS E DEFINIÇÃO DAS TAREFAS

Aprendizado de máquina designa métodos que ajustam parâmetros a partir de dados para produzir previsões ou representações. No aprendizado supervisionado, o treinamento depende de exemplos associados a rótulos; no autossupervisionado, a representação é aprendida por tarefas construídas a partir dos próprios dados. Redes convolucionais exploram relações espaciais locais, enquanto mecanismos de atenção permitem modelar dependências entre regiões. A escolha do método deve responder à tarefa e às características dos dados, sem pressupor superioridade clínica de uma arquitetura pelo seu grau de complexidade.

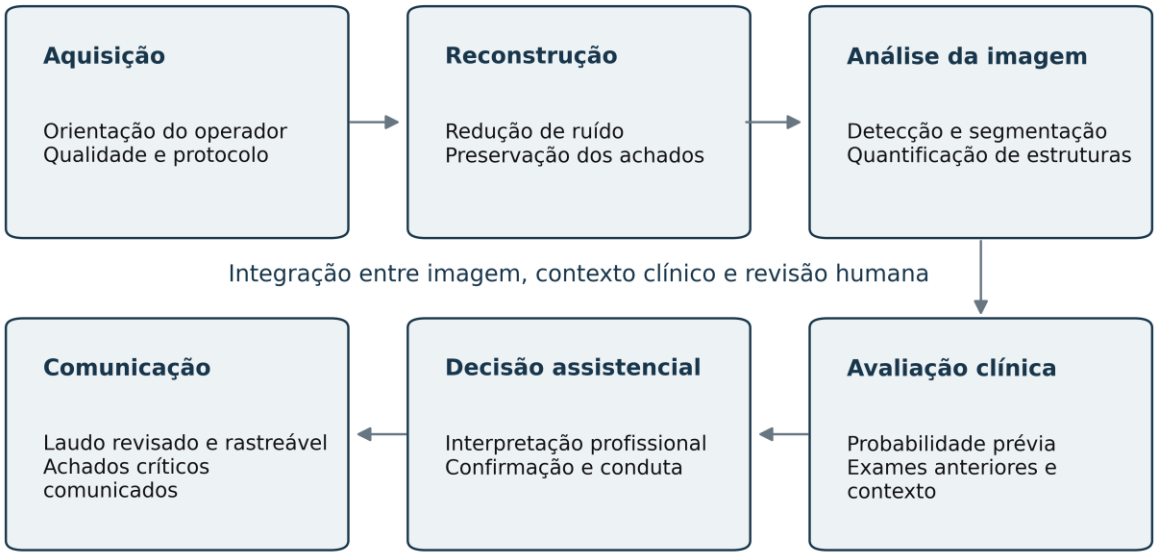
Classificação estima categorias ou probabilidades; detecção localiza achados; segmentação delimita estruturas em pixels ou voxels. Uma rede pode segmentar um nódulo com sobreposição elevada em relação à anotação sem distinguir sua etiologia. Da mesma forma, detectar uma opacidade não confirma pneumonia, pois o significado do achado depende de sintomas, evolução e diagnósticos diferenciais. A separação entre essas tarefas evita transformar uma saída matemática em conclusão clínica não sustentada.

A U-Net organiza a análise por um caminho de redução da resolução, que captura contexto, e outro de recuperação espacial, conectado às representações anteriores. O nnU-Net automatiza decisões de configuração para diferentes conjuntos de segmentação. Mais recentemente, o MedSAM investigou segmentação orientada por indicações do usuário em múltiplas modalidades. A amplitude do treinamento desses sistemas é relevante, mas não comprova adequação de todos os contornos a decisões terapêuticas ou

dispensa revisão da estrutura segmentada (Ronneberger; Fischer; Brox, 2015; Isensee et al., 2021; Ma et al., 2024).

Modelos de linguagem e sistemas multimodais acrescentam a capacidade de relacionar texto e imagem ou organizar documentos. Entretanto, produzir uma resposta linguisticamente coerente não demonstra que todas as afirmações estão presentes no exame. A finalidade de uso deve distinguir revisão textual, síntese de achados fornecidos e interpretação visual, porque essas atividades possuem fontes de erro e requisitos de validação diferentes (World Health Organization, 2024).

Figura 1 – Pontos de atuação da inteligência artificial no processo diagnóstico



Fonte: elaboração própria. Esquema conceitual de integração das tarefas, sem dados de pacientes. As setas representam relações de trabalho; a interpretação clínica integra as informações da imagem e do paciente.

3 APLICAÇÕES CLÍNICAS POR MODALIDADE

3.1 MAMOGRAFIA E RASTREAMENTO DO CÂNCER DE MAMA

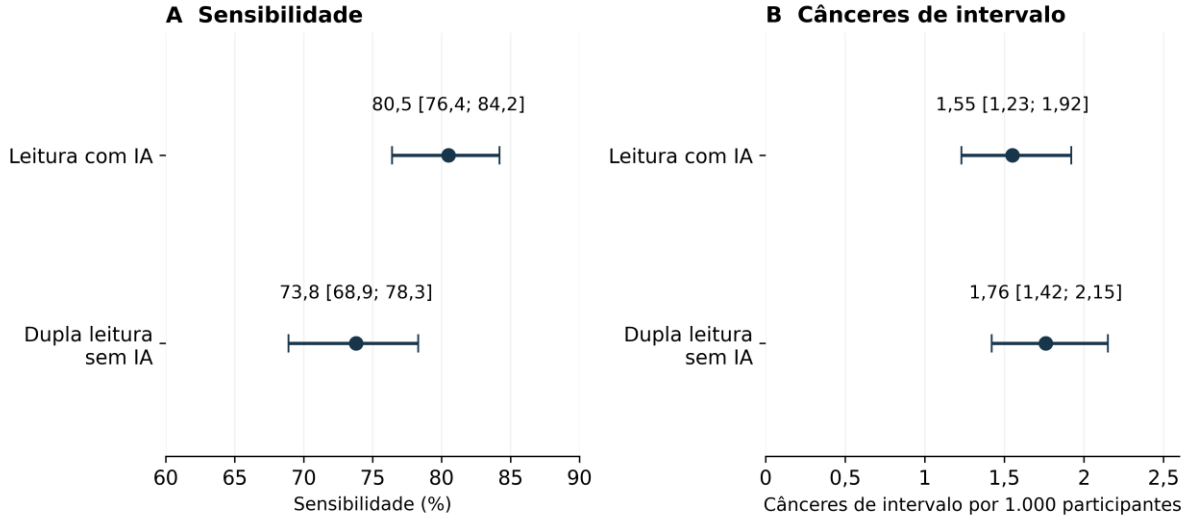
Na mamografia, a IA pode funcionar como apoio à detecção e à seleção de exames para diferentes regimes de leitura. McKinney et al. (2020) apresentaram avaliação retrospectiva internacional de um sistema para rastreamento, mas simulações de carga de trabalho não equivalem, por si, à implantação prospectiva. A contribuição desse tipo de estudo é estabelecer plausibilidade e desempenho em conjuntos definidos; a comprovação do efeito no serviço exige acompanhar o processo de leitura e seus resultados.

O ensaio sueco MASAI avançou nessa direção ao comparar um fluxo com IA e a dupla leitura convencional. Na publicação de 2025, a detecção foi de 6,4 versus 5,0 cânceres por 1.000 participantes, com redução aproximada de 44% da carga de leitura. Esse desfecho representa trabalho de interpretação

dentro do protocolo e não redução equivalente de toda a atividade do serviço. Os resultados se referem ao sistema, à versão e à organização da leitura estudados (Hernström et al., 2025).

Na análise de 2026, a taxa de câncer de intervalo foi de 1,55 versus 1,76 por 1.000 participantes, com razão de 0,88 e IC 95% de 0,65 a 1,18. O estudo demonstrou não inferioridade segundo a margem predefinida; o intervalo não comprova superioridade para esse desfecho. A sensibilidade foi de 80,5% com IA e 73,8% no controle, com especificidade de 98,5% nos dois grupos. A Figura 2 mantém as unidades e os intervalos publicados, sem combinar desfechos distintos (Gommers et al., 2026).

Figura 2 – Sensibilidade e cânceres de intervalo no ensaio MASAI



Fonte: elaboração própria a partir de Gommers et al. (2026). Pontos: estimativas publicadas; barras: IC 95%. A: porcentagem de sensibilidade. B: cânceres de intervalo por 1.000 participantes. São desfechos do mesmo ensaio, não estudos independentes. A não inferioridade da taxa de câncer de intervalo não demonstra superioridade nem redução de mortalidade.

3.2 RADIOGRAFIA TORÁCICA E PRIORIZAÇÃO DE EXAMES

Na radiografia de tórax, tarefas frequentes incluem detecção de opacidades, derrame pleural, pneumotórax e nódulos. O uso pretendido pode ser apoio à interpretação ou ordenação da fila de trabalho. Essas finalidades precisam ser avaliadas separadamente: um sistema que identifica um achado não necessariamente reduz o tempo até sua confirmação, e a priorização de alguns exames pode deslocar outros para posições posteriores.

O LungIMPACT avaliou prospectivamente a priorização de radiografias solicitadas na atenção primária em cinco serviços do sistema britânico. O ensaio não demonstrou redução significativa do tempo até a tomografia ou até o diagnóstico de câncer de pulmão. O resultado é relevante porque testa uma intervenção operacional, e não apenas acurácia retrospectiva. A identificação algorítmica de suspeita não supera, isoladamente, barreiras de encaminhamento, agendamento e investigação subsequente (Woznitza et al., 2026).

A generalização também merece atenção. Zech et al. (2018) demonstraram variação do desempenho de um modelo de pneumonia entre instituições e a possibilidade de explorar características associadas à origem dos exames. Assim, resultados internos elevados podem refletir diferenças de equipamento, prevalência ou processo de aquisição, além dos sinais da doença. A validação local deve incluir exames consecutivos e condições de aquisição semelhantes às do uso pretendido.

3.3 TOMOGRAFIA COMPUTADORIZADA E ONCOLOGIA TORÁCICA

Na tomografia, o processamento volumétrico permite localizar lesões, medir estruturas e comparar exames seriados. Ardila et al. (2019) investigaram um sistema tridimensional para rastreamento de câncer de pulmão em tomografia de baixa dose, incluindo validação independente. O estudo oferece evidência de capacidade discriminativa em um contexto delimitado de rastreamento, sem autorizar extrapolação para todas as tomografias, populações sintomáticas ou protocolos contrastados.

Na aplicação clínica, é necessário distinguir detecção de nódulo, estimativa de malignidade e recomendação de seguimento. Essas saídas não são intercambiáveis: a decisão depende do tamanho, da evolução, de fatores de risco e da informação disponível no exame. Comparações longitudinais ainda exigem atenção a espessura de corte, reconstrução e segmentação, pois variações técnicas podem produzir aparente mudança de volume sem alteração biológica correspondente.

3.4 RESSONÂNCIA MAGNÉTICA E RECONSTRUÇÃO ACELERADA

Na ressonância magnética, uma aplicação relevante é reconstruir imagens a partir de aquisições subamostradas. Johnson et al. (2023) avaliaram prospectivamente 170 participantes em ressonância de joelho, com interpretação por seis radiologistas musculoesqueléticos. O protocolo com reconstrução por aprendizado profundo permitiu redução próxima de duas vezes no tempo de aquisição, com equivalência diagnóstica no contexto investigado. Esse resultado não equivale a validar automaticamente a técnica para todos os órgãos ou achados.

A qualidade visual deve ser diferenciada da fidelidade diagnóstica. Reduzir ruído pode facilitar a leitura, mas a reconstrução precisa preservar lesões pequenas, bordas e contrastes relevantes. Antun et al. (2020) demonstraram instabilidades de métodos de reconstrução baseados em aprendizado profundo, incluindo sensibilidade a perturbações e a alterações na imagem. A avaliação clínica deve investigar falhas plausíveis e não se limitar a medidas médias de similaridade ou preferência estética.

3.5 ULTRASSONOGRAFIA E ECOCARDIOGRAFIA

Na ultrassonografia, a IA pode orientar a aquisição antes da interpretação, identificando planos e sugerindo ajustes do transdutor. Narang et al. (2021) estudaram o uso de orientação por aprendizado

profundo para aquisição de ecocardiogramas limitados por profissionais sem experiência prévia em ultrassonografia. A possibilidade de obter imagens adequadas para parâmetros específicos é distinta da realização autônoma de um exame completo.

A aplicação exige delimitar quais estruturas e perguntas clínicas podem ser avaliadas, quais imagens são insuficientes e quando encaminhar para exame especializado. Obesidade, janela acústica desfavorável, movimento e condições anatômicas particulares podem dificultar a aquisição. Portanto, ampliar acesso deve vir acompanhado de critérios de qualidade e supervisão, evitando que a disponibilidade de uma imagem seja confundida com exclusão de todas as doenças cardíacas.

Quadro 1 – Aplicações selecionadas e limites de extrapolação

Modalidade ou tarefa	**Contribuição avaliada**	**Limite da interpretação**
Mamografia	Apoio à leitura e redistribuição do trabalho no MASAI.	Resultados vinculados ao programa, à versão e ao protocolo estudados.
Radiografia torácica	Priorização na investigação de câncer no LungIMPACT.	A priorização isolada não demonstrou acelerar o diagnóstico.
Tomografia de baixa dose	Modelagem tridimensional do risco no rastreamento pulmonar.	Desempenho de rastreamento não se aplica a todo exame torácico.
Ressonância de joelho	Reconstrução acelerada avaliada prospectivamente.	Equivalência no protocolo estudado não garante preservação de todo achado.
Ecocardiografia limitada	Orientação de aquisição por operadores iniciantes.	Adequação para parâmetros selecionados não equivale a exame completo.

Fonte: síntese de Hernström et al. (2025), Gommers et al. (2026), Woznitza et al. (2026), Ardila et al. (2019), Johnson et al. (2023) e Narang et al. (2021).

4 QUANTIFICAÇÃO RADIÔMICA E MODELOS DE LINGUAGEM

A radiômica transforma regiões de interesse em descritores quantitativos de intensidade, forma e textura. Diferentemente de uma medida convencional isolada, esses conjuntos de atributos podem apoiar modelos de associação ou predição. Contudo, a repetibilidade depende da aquisição, da reconstrução, da segmentação e das definições matemáticas. A Image Biomarker Standardization Initiative estabeleceu definições e referências para tornar a extração mais comparável entre implementações (Zwanenburg et al., 2020).

Um modelo radiômico pode apresentar associação com prognóstico sem acrescentar informação útil a variáveis clínicas já disponíveis. Para demonstrar valor incremental, deve ser comparado a um modelo de referência, em população independente, com análise de calibração e da decisão que se pretende modificar. Selecionar atributos antes de separar os conjuntos ou repetir escolhas com base no teste aumenta o risco de desempenho otimista. A validade exige avaliar todo o processo, não apenas o classificador final.

No processamento de laudos, modelos de linguagem podem identificar inconsistências, organizar achados e produzir versões de comunicação. Gertz et al. (2024) avaliaram a detecção de erros em laudos, com resultados comparáveis aos dos radiologistas naquele desenho experimental. O objeto do estudo era a

identificação de erros textuais; isso não valida a geração autônoma de diagnóstico a partir da imagem, nem estabelece ausência de falhas na rotina.

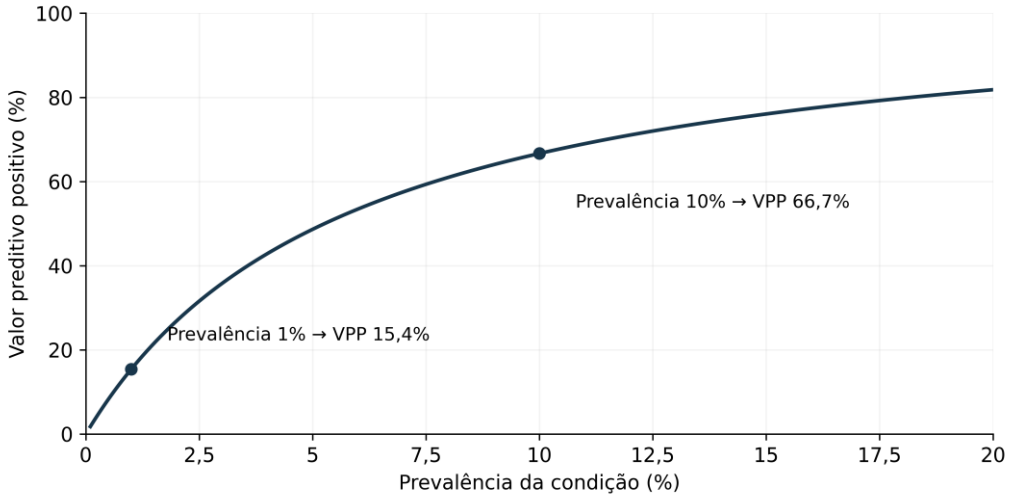
Uma saída textual pode omitir negação, lateralidade, medida ou recomendação, alterando o significado clínico. A revisão precisa confrontar a versão gerada com os achados originais e com o exame correspondente. Modelos multimodais demandam, adicionalmente, avaliação da correspondência entre o que foi observado e o que foi descrito. A governança de sistemas generativos deve considerar respostas factualmente incorretas, vieses e exposição de dados, além da fluência do texto (World Health Organization, 2024).

5 COMO INTERPRETAR O DESEMPENHO DIAGNÓSTICO

A área sob a curva ROC resume a discriminação ao variar o limiar de decisão. Ela não determina o desempenho no limiar adotado nem informa, isoladamente, se uma probabilidade está calibrada. Sensibilidade e especificidade devem ser apresentadas com seus intervalos de confiança e com o padrão de referência utilizado. Valores preditivos dependem também da prevalência da condição na população avaliada, aspecto particularmente relevante ao transferir sistemas entre rastreamento e serviços de alta suspeição.

A Figura 3 apresenta uma simulação didática: com sensibilidade de 90% e especificidade de 95% mantidas constantes, o valor preditivo positivo é de aproximadamente 15,4% quando a prevalência é de 1%, chegando a 66,7% com prevalência de 10%. Os números derivam do teorema de Bayes, não de uma ferramenta clínica. O exemplo mostra por que uma grande proporção de alertas pode ser falsa em populações de baixa prevalência, apesar de medidas aparentemente favoráveis.

Figura 3 – Efeito da prevalência sobre o valor preditivo positivo



Fonte: elaboração própria. Simulação matemática, sem dados de pacientes ou estimativa de desempenho de um produto. Sensibilidade fixa de 90% e especificidade fixa de 95%. VPP = sensibilidade × prevalência / [sensibilidade × prevalência + (1 − especificidade) × (1 − prevalência)].

Calibração avalia a concordância entre probabilidades previstas e frequências observadas. Para uma previsão de risco orientar investigação, uma estimativa de 10% deve possuir interpretação coerente na população de uso. Já o benefício líquido incorpora o peso relativo de falsos positivos e falsos negativos em limiares clinicamente relevantes. A análise de curvas de decisão permite comparar modelos e estratégias, mas depende de limiares justificados e não substitui avaliação prospectiva do cuidado (Vickers; Elkin, 2006).

Na segmentação, medidas de sobreposição, como Dice, precisam ser acompanhadas da localização dos erros. Uma pequena falha de contorno pode ter importância desproporcional quando próxima a uma estrutura crítica, enquanto uma diferença volumétrica semelhante em outra região pode não alterar a conduta. O critério de sucesso deve refletir a finalidade: mensuração, planejamento, acompanhamento ou apoio à detecção. Essa definição precisa anteceder a escolha da métrica.

6 LIMITAÇÕES E FONTES DE ERRO

6.1 DADOS DE REFERÊNCIA E VAZAMENTO DE INFORMAÇÃO

Rótulos obtidos de laudos podem reproduzir incertezas e práticas locais, enquanto consenso entre especialistas não constitui padrão infalível para toda condição. O estudo deve explicar quem anotou as imagens, quais informações estavam disponíveis e como divergências foram resolvidas. A ausência de doença pode exigir seguimento ou outro procedimento de confirmação, dependendo da pergunta clínica. O CLAIM orienta o relato transparente desses elementos e da composição dos conjuntos de dados (Tejani et al., 2024).

O vazamento ocorre quando informações do conjunto de avaliação influenciam o treinamento ou as escolhas do modelo. Cortes da mesma tomografia, imagens da mesma pessoa ou exames seriados distribuídos entre treino e teste podem tornar o problema artificialmente fácil. A separação deve respeitar o paciente e a estrutura temporal; o conjunto de teste não deve orientar seleção de atributos, ajuste de hiperparâmetros ou escolha do limiar. A descrição explícita dessas decisões é necessária para interpretar a validação (Tejani et al., 2024).

6.2 MUDANÇA DE DOMÍNIO E DESIGUALDADES

A mudança de domínio envolve diferenças entre os dados de desenvolvimento e os de utilização, como equipamento, protocolo, população e prevalência. A distinção entre validação interna e externa é, portanto, substantiva. O estudo de Zech et al. (2018) ilustra por que um algoritmo pode explorar sinais ligados ao hospital e perder desempenho em outra instituição. Mesmo uma avaliação multicêntrica não cobre necessariamente todos os cenários em que o sistema será utilizado.

A média populacional também pode ocultar subdiagnóstico em grupos específicos. Seyyed-Kalantari et al. (2021) identificaram desigualdades de subdiagnóstico em modelos aplicados a radiografias de tórax de populações pouco atendidas. Auditar sensibilidade e taxas de erro por subgrupo, com amostras e incerteza adequadas, é necessário para reconhecer danos que não aparecem no indicador agregado. A ausência de diferença estatística em subgrupos pequenos não comprova equidade.

6.3 INTERAÇÃO ENTRE PROFISSIONAL E SISTEMA

O desempenho do conjunto profissional–IA não pode ser deduzido da comparação entre profissional e algoritmo isolados. O alerta pode chamar atenção para um achado verdadeiro, mas também induzir aceitação de uma sugestão incorreta ou desviar a busca de outras alterações. Estudos de leitura devem considerar ordem de apresentação, treinamento, tempo, experiência dos participantes e distribuição dos casos. O DECIDE-AI destaca a avaliação inicial em ambiente clínico e os fatores humanos envolvidos na implementação (Vasey et al., 2022).

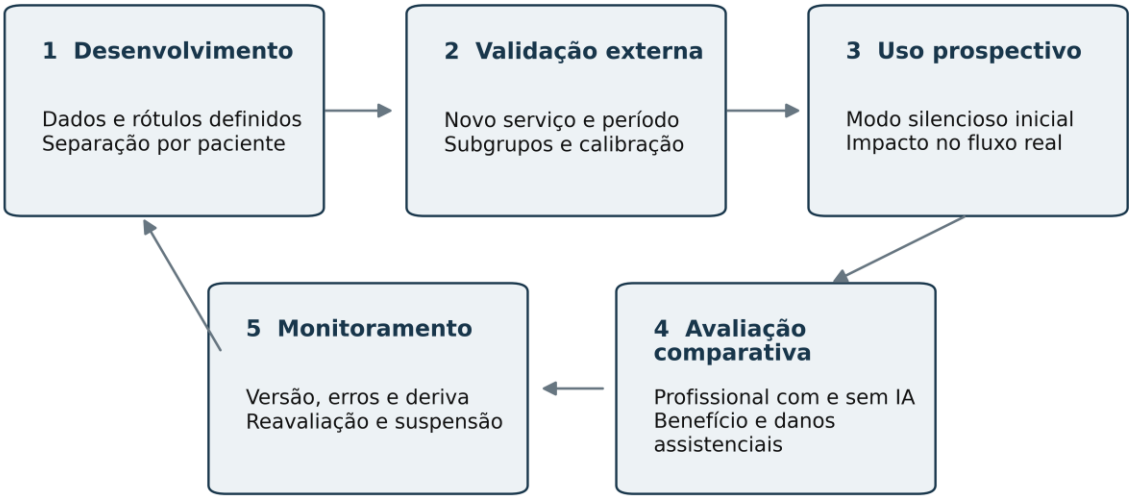
Mapas de atenção e regiões destacadas podem auxiliar inspeção, mas não demonstram que o algoritmo utilizou uma relação causal válida. A interpretação clínica deve manter a possibilidade de discordância fundamentada, com registro das falhas. Uma interface útil precisa comunicar incerteza e limitações sem converter a probabilidade computacional em autoridade indevida sobre o caso.

7 IMPLEMENTAÇÃO E MONITORAMENTO NO SERVIÇO

A implantação deve começar por uma pergunta operacional: qual decisão será modificada e qual resultado justificará a mudança? Uma ferramenta de priorização precisa demonstrar efeito na fila e nos tempos relevantes; uma ferramenta de quantificação deve demonstrar consistência e utilidade da medida. O estudo inicial em modo silencioso permite observar as saídas sem alterar o atendimento, desde que sua finalidade e seus limites sejam previamente definidos. A avaliação posterior deve acompanhar o sistema em uso, conforme a proposta do DECIDE-AI (Vasey et al., 2022).

A integração ao sistema de arquivamento e comunicação de imagens (PACS) e ao sistema de informação radiológica deve preservar identificação do exame, lateralidade, versão e disponibilidade das séries utilizadas. O serviço precisa definir quem revisa as saídas, como se comunica um achado crítico e o que ocorre em caso de falha ou indisponibilidade. O monitoramento deve contemplar erros, proporção de exames processados, tempos, distribuição das saídas e mudanças do perfil de pacientes. Atualizações de software ou de protocolo podem exigir nova avaliação.

Figura 4 – Etapas de avaliação e monitoramento de um sistema de IA em imagem



Fonte: elaboração própria, em síntese conceitual de Tejani et al. (2024), Vasey et al. (2022) e World Health Organization (2021). As etapas organizam a avaliação; não representam certificação regulatória nem garantem ausência de eventos adversos.

A avaliação econômica inclui mais do que o preço da licença: integração, processamento, treinamento, revisão de alertas, manutenção e investigação adicional também consomem recursos. Ganhar tempo em uma tarefa não implica reduzir custo total se a intervenção aumentar exames desnecessários ou transferir trabalho para outra equipe. O LungIMPACT reforça a necessidade de avaliar a trajetória assistencial completa, porque priorizar uma imagem não assegura acelerar todas as etapas seguintes (Woznitza et al., 2026).

Quadro 2 – Perguntas essenciais para avaliar a adoção no serviço

Dimensão	Pergunta que precisa ser respondida
Uso pretendido	Qual tarefa, população e decisão serão modificadas?
Validade	Houve avaliação independente, com padrão de referência e denominadores claros?
Generalização	O protocolo, o equipamento e a população locais estão representados?
Benefício e danos	O sistema melhora uma decisão ou um resultado relevante, sem dano desproporcional?
Integração	Há revisão, comunicação de achados críticos e plano para indisponibilidade?
Monitoramento	Versões, falhas, mudanças de desempenho e critérios de suspensão serão registrados?

Fonte: elaboração própria, a partir de Tejani et al. (2024), Vasey et al. (2022) e World Health Organization (2021).

8 ÉTICA E GOVERNANÇA NA PRÁTICA MÉDICA

A governança envolve transparência sobre a finalidade, proteção dos dados, responsabilização e participação dos profissionais e das pessoas afetadas. A orientação da Organização Mundial da Saúde situa autonomia, segurança, equidade e interesse público no desenvolvimento e na utilização da IA. O serviço deve compatibilizar acesso aos dados com sua finalidade, reduzir exposição desnecessária e manter

mecanismos de identificação e resposta a danos, em vez de transferir toda a responsabilidade ao operador final (World Health Organization, 2021).

As perspectivas incluem modelos capazes de aproveitar representações aprendidas em grande escala e sistemas que relacionem imagem, laudo e contexto clínico. O MedSAM exemplifica a adaptação de modelos de segmentação a diferentes modalidades, mas a flexibilidade depende das indicações fornecidas e das condições de avaliação (Ma et al., 2024). A expansão das tarefas deve vir acompanhada de testes específicos, pois a competência em um domínio não estabelece confiabilidade em outro.

Para os modelos generativos, a prioridade clínica é preservar fatos, rastreabilidade e delimitação do uso. Respostas confiantes podem conter afirmações não sustentadas, e a combinação de múltiplas fontes amplia a necessidade de verificar de onde cada informação foi obtida. As recomendações da OMS para modelos multimodais enfatizam a governança dessas aplicações e seus riscos. A avaliação deve acompanhar mudanças de versão e de contexto, sem pressupor desempenho estável apenas porque a interface permanece semelhante (World Health Organization, 2024).

Na formação médica, torna-se necessário compreender métricas, prevalência, incerteza e limites de extrapolação. A contribuição do profissional inclui formular a pergunta clínica, reconhecer informação ausente, integrar achados discordantes e decidir quando uma saída automatizada não se aplica. O desenvolvimento da área deve ser acompanhado por estudos que avaliem essas interações, a carga real de trabalho e consequências para pacientes, em diferentes sistemas de saúde.

9 CONSIDERAÇÕES FINAIS

A inteligência artificial já oferece contribuições mensuráveis em tarefas delimitadas do diagnóstico por imagem, como organização da leitura mamográfica, reconstrução acelerada, segmentação e revisão de laudos. Entretanto, acurácia e produtividade não garantem, isoladamente, benefício clínico. A evidência deve ser interpretada segundo a finalidade, a população, o protocolo e a organização do serviço. Validação externa, avaliação prospectiva, análise de subgrupos e monitoramento são componentes do processo de adoção. A incorporação responsável depende de integrar as saídas computacionais ao julgamento clínico, mantendo rastreabilidade, proteção dos dados e capacidade de reconhecer e corrigir falhas.

REFERÊNCIAS

ANTUN, V. et al. On instabilities of deep learning in image reconstruction and the potential costs of AI. **Proceedings of the National Academy of Sciences**, v. 117, n. 48, p. 30088–30095, 2020. DOI: 10.1073/pnas.1907377117.

ARDILA, D. et al. End-to-end lung cancer screening with three-dimensional deep learning on low-dose chest computed tomography. **Nature Medicine**, v. 25, p. 954–961, 2019. DOI: 10.1038/s41591-019-0447-x.

GERTZ, R. J. et al. Potential of GPT-4 for detecting errors in radiology reports: implications for reporting accuracy. **Radiology**, v. 311, n. 1, e232714, 2024. DOI: 10.1148/radiol.232714.

GOMMERS, J. et al. Interval cancer, sensitivity, and specificity comparing AI-supported mammography screening with standard double reading without AI in the MASAI study: a randomised, controlled, non-inferiority, single-blinded, population-based, screening-accuracy trial. **The Lancet**, v. 407, n. 10527, p. 505–514, 2026. DOI: 10.1016/S0140-6736(25)02464-X.

HERNSTRÖM, V. et al. Screening performance and characteristics of breast cancer detected in the Mammography Screening with Artificial Intelligence trial (MASAI): a randomised, controlled, parallel-group, non-inferiority, single-blinded, screening accuracy study. **The Lancet Digital Health**, v. 7, n. 3, p. e175–e183, 2025. DOI: 10.1016/S2589-7500(24)00267-X.

ISENSEE, F. et al. nnU-Net: a self-configuring method for deep learning-based biomedical image segmentation. **Nature Methods**, v. 18, p. 203–211, 2021. DOI: 10.1038/s41592-020-01008-z.

JOHNSON, P. M. et al. Deep learning reconstruction enables prospectively accelerated clinical knee MRI. **Radiology**, v. 307, n. 2, e220425, 2023. DOI: 10.1148/radiol.220425.

MA, J. et al. Segment anything in medical images. **Nature Communications**, v. 15, art. 654, 2024. DOI: 10.1038/s41467-024-44824-z.

MCKINNEY, S. M. et al. International evaluation of an AI system for breast cancer screening. **Nature**, v. 577, p. 89–94, 2020. DOI: 10.1038/s41586-019-1799-6.

NARANG, A. et al. Utility of a deep-learning algorithm to guide novices to acquire echocardiograms for limited diagnostic use. **JAMA Cardiology**, v. 6, n. 6, p. 624–632, 2021. DOI: 10.1001/jamacardio.2021.0185.

RONNEBERGER, O.; FISCHER, P.; BROX, T. U-Net: convolutional networks for biomedical image segmentation. In: NAVAB, N. et al. (ed.). **Medical Image Computing and Computer-Assisted Intervention – MICCAI 2015**. Cham: Springer, 2015. p. 234–241. DOI: 10.1007/978-3-319-24574-4_28.

SEYYED-KALANTARI, L. et al. Underdiagnosis bias of artificial intelligence algorithms applied to chest radiographs in under-served patient populations. **Nature Medicine**, v. 27, p. 2176–2182, 2021. DOI: 10.1038/s41591-021-01595-0.

TEJANI, A. S. et al. Checklist for Artificial Intelligence in Medical Imaging (CLAIM): 2024 update. **Radiology: Artificial Intelligence**, v. 6, n. 4, e240300, 2024. DOI: 10.1148/ryai.240300.

VASEY, B. et al. Reporting guideline for the early-stage clinical evaluation of decision support systems driven by artificial intelligence: DECIDE-AI. **Nature Medicine**, v. 28, p. 924–933, 2022. DOI: 10.1038/s41591-022-01772-9.

VICKERS, A. J.; ELKIN, E. B. Decision curve analysis: a novel method for evaluating prediction models. **Medical Decision Making**, v. 26, n. 6, p. 565–574, 2006. DOI: 10.1177/0272989X06295361.

WORLD HEALTH ORGANIZATION. Ethics and governance of artificial intelligence for health. Geneva: **WHO**, 2021. Disponível em: https://www.who.int/publications/i/item/9789240029200. Acesso em: 26 out. 2026.

WORLD HEALTH ORGANIZATION. Ethics and governance of artificial intelligence for health: guidance on large multi-modal models. Geneva: **WHO**, 2024. Disponível em: https://www.who.int/publications/i/item/9789240084759. Acesso em: 26 out. 2026.

WOZNITZA, N. et al. AI-based chest X-ray prioritization in the lung cancer diagnostic pathway: the LungIMPACT randomized controlled trial. **Nature Medicine**, v. 32, n. 5, p. 1737–1744, 2026. DOI: 10.1038/s41591-026-04253-5.

ZECH, J. R. et al. Variable generalization performance of a deep learning model to detect pneumonia in chest radiographs: a cross-sectional study. **PLOS Medicine**, v. 15, n. 11, e1002683, 2018. DOI: 10.1371/journal.pmed.1002683.

ZWANENBURG, A. et al. The Image Biomarker Standardization Initiative: standardized quantitative radiomics for high-throughput image-based phenotyping. **Radiology**, v. 295, n. 2, p. 328–338, 2020. DOI: 10.1148/radiol.2020191145.