

PODEMOS CONFIAR NA INTELIGÊNCIA ARTIFICIAL? EXPLICABILIDADE, ERROS E SEGURANÇA NO DIAGNÓSTICO POR IMAGEM

CAN WE TRUST ARTIFICIAL INTELLIGENCE? EXPLAINABILITY, ERRORS, AND SAFETY IN MEDICAL IMAGING DIAGNOSIS

Clarice Rodrigues de Almeida¹, Raphaela Maria Costa Pinto de Oliveira¹, Katarinne Maria Rocha Moura Beer¹, Andréia Galvão Schelb¹, Victor Menezes Cardoso², Taís Silva Souza¹.

1. Acadêmico do curso de Medicina do Centro Universitário de Maceió UNIMA-Afya.
2. Professor do curso de Medicina do Centro Universitário de Maceió UNIMA-Afya.
clarice.almeida.med@gmail.com

EIXO: Tecnologia e inovação

RESUMO

Introdução: A inteligência artificial transformou o diagnóstico por imagem, mas sua transição para a prática clínica enfrenta desafios de validação. **Objetivo:** Analisar as evidências científicas atuais sobre a explicabilidade, erros diagnósticos e segurança desses sistemas, identificando fatores determinantes para sua confiabilidade clínica. **Método:** Revisão integrativa baseada no protocolo PRISMA, realizada nas bases PubMed, Scopus, Web of Science, Biblioteca Virtual em Saúde e MedLine, com artigos de 2024 a 2026. A amostra final compreendeu sete artigos. **Resultados:** Os achados indicam que a alta acurácia teórica não garante confiabilidade imediata. Erros algorítmicos são imprevisíveis e os métodos atuais de explicabilidade falham em demonstrar o nexo causal clínico, podendo induzir o médico a falhas. **Conclusão:** A segurança do paciente depende de um modelo de trabalho complementar, posicionando o software como suporte de segunda leitura e mantendo o radiologista como a autoridade soberana e responsável legal pelo laudo médico.

Palavras-chave: Inteligência Artificial. Erros Diagnósticos. Diagnóstico por Imagem. Radiologia. Segurança do Paciente.

ABSTRACT

Introduction: Artificial intelligence has transformed diagnostic imaging, but its transition into clinical practice faces validation challenges. **Objective:** To analyze current scientific evidence on the explainability, diagnostic errors, and safety of these systems, identifying determining factors for their clinical reliability. **Method:** Integrative review based on the PRISMA protocol, conducted in the PubMed, Scopus, Web of Science, Virtual Health Library, and MedLine databases, using articles from 2024 to 2026. The final sample comprised seven papers. **Results:** The findings indicate that high theoretical accuracy does not guarantee immediate reliability. Algorithmic errors are unpredictable and current explainability methods fail to demonstrate clinical causal links, potentially inducing physician error. **Conclusion:** Patient safety depends on a complementary working model,

positioning the software as secondary reading support and keeping the human radiologist as the sovereign authority and legally responsible party for the medical report.

Keywords: Artificial Intelligence. Diagnostic Errors. Diagnostic Imaging. Radiology. Patient Safety.

INTRODUÇÃO

A evolução dos algoritmos de aprendizado transformou a medicina ao estabelecer novos horizontes na análise de exames complexos¹. Essa rápida ascensão tecnológica carrega uma relevância social e científica, dado o potencial de otimizar fluxos de trabalho hospitalares, reduzir gargalos assistenciais e levar laudos de alta qualidade a regiões desprovidas de especialistas de nicho. Contudo, a transição desses modelos de desenvolvimento impõe desafios de validação analítica e segurança operacional².

Savage e colaboradores³ apontam que a ausência de mecanismos transparentes de interpretabilidade gera uma lacuna de confiança crônica entre os profissionais de saúde e a tecnologia, dificultando a validação clínica imediata dos laudos gerados. A incapacidade de discernir com exatidão quais regiões anatômicas ou características de pixel influenciaram a decisão da máquina impede que o radiologista identifique vieses e falhas operacionais estruturais. Consequentemente, estudos recentes indicam que altos índices de acurácia estatística reportados em laboratório não se traduzem necessariamente em alta confiabilidade à beira do leito⁴.

Essa assimetria entre o desempenho teórico e a prática médica expõe uma vulnerabilidade crítica relacionada à natureza dos erros algorítmicos. Investigações empíricas demonstram que modelos de inteligência artificial falham de maneira imprevisível quando expostos a variações de artefatos técnicos, diferentes marcas de tomógrafos ou patologias raras não previstas no treinamento, induzindo a erros sistemáticos de falsos positivos e falsos negativos⁵. Diante desse cenário, emerge a necessidade premente de estabelecer critérios rigorosos de segurança e protocolos robustos de supervisão humana. A literatura atual defende que o futuro ideal da especialidade não reside na substituição do médico, mas sim em um modelo de trabalho simbiótico e complementar, onde a experiência clínica humana atue como a linha final de defesa para mitigar riscos e assegurar a integridade do paciente⁶.

O avanço seguro da radiologia digital depende diretamente do mapeamento dos gargalos éticos, técnicos e operacionais que ainda impedem a adoção plena de sistemas inteligentes na rotina médica diagnóstica. O objetivo desta pesquisa é analisar as evidências

científicas atuais sobre a explicabilidade, os erros diagnósticos e a segurança dos sistemas de inteligência artificial aplicados ao diagnóstico por imagem, identificando os fatores determinantes para a sua confiabilidade na prática clínica e discutindo a importância da supervisão do radiologista na salvaguarda da segurança do paciente.

MÉTODO

O presente estudo caracteriza-se como uma revisão integrativa da literatura, de caráter descritivo e analítico, desenvolvida para sintetizar as evidências científicas atuais sobre a inteligência artificial aplicada à medicina. O processo de seleção e análise dos dados seguiu os princípios metodológicos do protocolo PRISMA para garantir a transparência da busca estruturada. Esta ocorreu nas bases de dados PubMed, Scopus, Web of Science, Biblioteca Virtual em Saúde (BVS) e MedLine, delimitada ao período de 2024 a 2026. Combinaram-se descritores controlados e termos livres via operadores booleanos, correlacionando os termos inteligência artificial, erros diagnósticos, segurança, IA explicável e radiologia.

A condução de todo o processo de triagem e análise foi direcionada de forma a responder detalhadamente à pergunta norteadora deste estudo: "Quais são as evidências atuais sobre explicabilidade, erros diagnósticos e segurança dos sistemas de inteligência artificial aplicados ao diagnóstico por imagem, e quais fatores determinam sua confiabilidade na prática clínica?". Os critérios de inclusão selecionaram artigos empíricos, revisões sistemáticas, de escopo ou ensaios clínicos publicados entre 2024 e 2026 focados em desempenho diagnóstico, erros como falsos positivos ou negativos, vieses algorítmicos, mecanismos de explicabilidade, segurança do paciente e supervisão humana no trabalho complementar entre inteligência artificial e o radiologista, nos idiomas inglês, português ou espanhol. Foram excluídos estudos anteriores a 2024, pesquisas puramente técnicas de engenharia de software sem aplicação clínica prática, editoriais, cartas ao editor, resumos de congressos, relatórios comerciais e registros duplicados.

A triagem bibliográfica obedeceu a um fluxo rigoroso de seleção. Inicialmente, identificaram-se 115 artigos nas bases de dados avaliadas, sendo 35 no PubMed, 30 no Scopus, 25 no Web of Science, 15 no MedLine e 10 na BVS. Foram filtrados e removidos 23 artigos em duplicidade, restando 92 produções para avaliação preliminar de título e resumo. Após essa etapa, 68 artigos foram excluídos por desvio evidente do escopo clínico estabelecido. Dos 24 artigos selecionados para leitura integral, 18 foram descartados devido a restrições metodológicas específicas, como foco exclusivo em modelagem matemática e

engenharia computacional pura ou ausência crônica de dados aplicados à segurança do paciente. Resultando numa amostra final de 6 artigos de alta relevância científica, com dados extraídos e sintetizados de forma contínua em torno dos pilares conceituais de explicabilidade, erros, segurança e supervisão do radiologista, confrontando diretamente a acurácia teórica dos modelos com a real confiabilidade deles no cenário prático da rotina médica.

RESULTADOS

A análise detalhada dos 6 artigos científicos selecionados revelou um panorama convergente sobre as potencialidades e as vulnerabilidades da aplicação de sistemas inteligentes na rotina diagnóstica. Para sintetizar os achados identificados na literatura de forma lógica e integrada, as informações foram agrupadas com base nos três pilares temáticos do estudo: mecanismos de explicabilidade algorítmica, caracterização e origem dos erros diagnósticos, e os requisitos fundamentais de segurança sob supervisão humana. A distribuição cronológica, as características metodológicas e as contribuições centrais de cada investigação bibliográfica estão consolidadas no Quadro 1, que mapeia as evidências contemporâneas sobre o comportamento da tecnologia na radiologia.

Quadro 1: *Síntese dos achados principais dos estudos selecionados sobre inteligência artificial no diagnóstico por imagem.*

Autor e Ano Objetivo do Estudo		Aspecto Avaliado	Principais Achados Clínicos
Do ¹ (2024)	Avaliar segurança operacional da IA na radiologia	Segurança	Exige métricas de confiança independentes do modelo para evitar excesso de confiança do médico.
Savage et al. ³ (2026)	Propor framework de transparência para imagens	Explicabilidade	Define que a explicabilidade útil precisa aliar robustez técnica ao contexto real do usuário final.
Li et al. ² (2025)	Analisar as causas de erros diagnósticos por IA	Erros	Divide as falhas em patologia de dados, viés algorítmico e problemas de interação humano-máquina.

Prinster et al. ⁵ (2024)	Mensurar o impacto de explicações no erro médico	Explicabilidade / Erros	Demonstra que mapas de calor e explicações incorretas induzem radiologistas a cometer erros graves.
Velamala et al. ⁶ (2025)	Avaliar redução de erros via escopo de revisão	Erros / Segurança	Mostra que a IA reduz falhas de fadiga humana, mas demanda rigorosa validação local antes do uso.
Singh et al. ⁴ (2025)	Investigar limitações de explicações <i>post hoc</i>	Explicabilidade	Adverte que mapas de saliência enganam o profissional e geram falsa sensação de segurança.

Fonte: Elaborado pelos autores (2026).

No que diz respeito ao pilar da explicabilidade, as evidências apontam que os métodos atuais de abertura da "caixa-preta", como os mapas de saliência e explicações *post hoc*, apresentam severas limitações práticas que comprometem a confiabilidade da tomada de decisão. Os dados indicam que tais ferramentas frequentemente falham em apontar o nexo causal real do raciocínio algorítmico, funcionando mais como um elemento de persuasão visual do que como um validador científico. No âmbito dos erros diagnósticos, a literatura documenta que as falhas operacionais não ocorrem de forma aleatória, estando intimamente ligadas a vieses de treinamento, variabilidade técnica de equipamentos e dificuldades de leitura de patologias raras. Por fim, quanto à segurança do paciente, os estudos enfatizam que a mitigação de eventos adversos depende da instituição de um modelo de trabalho complementar, onde o sistema inteligente atua como segunda leitura de triagem e o radiologista assume o papel definitivo de revisor técnico e legal do laudo médico.

CONCLUSÃO

A alta acurácia teórica da inteligência artificial não garante confiabilidade clínica, pois os erros algorítmicos são imprevisíveis e os métodos de explicabilidade atuais podem induzir o médico a falhas. Esta revisão demonstra que a segurança do paciente exige um modelo de trabalho complementar, posicionando o software como suporte de segunda leitura e o radiologista como responsável legal pelo laudo.

Pesquisas futuras devem desenvolver modelos explicáveis integrados ao raciocínio médico e validações locais focadas em desfechos clínicos reais. É necessário criar protocolos de treinamento para a identificação de vieses nos algoritmos. Essas medidas reduzem o excesso de confiança e garantem uma integração ética e segura na rotina assistencial.

REFERÊNCIAS

1. Do S. Explainable & Safe Artificial Intelligence in Radiology. *J Korean Soc Radiol*. 2024 Sep;85(5):834-847. doi: 10.3348/jksr.2024.0118. PMID: 39416324; PMCID: PMC11473981.
2. Savage CH, Sulam J, Huang CM, Yi PH. Explainable Artificial Intelligence (AI) for Medical Imaging: A Framework for Bridging the AI Trust Gap. *AJR Am J Roentgenol*. 2026 May 13. doi: 10.2214/AJR.26.34829. Epub ahead of print. PMID: 42126554.
3. Li Y, Yi X, Fu J, Yang Y, Duan C, Wang J. Reducing misdiagnosis in AI-driven medical diagnostics: a multidimensional framework for technical, ethical, and policy solutions. *Front Med (Lausanne)*. 2025 Oct 31;12:1594450. doi: 10.3389/fmed.2025.1594450. PMID: 41244777; PMCID: PMC12615213.
4. Prinster D, Mahmood A, Saria S, Jeudy J, Lin CT, Yi PH, Huang CM. Care to Explain? AI Explanation Types Differentially Impact Chest Radiograph Diagnostic Performance and Physician Trust in AI. *Radiology*. 2024 Nov;313(2):e233261. doi: 10.1148/radiol.233261. PMID: 39560483; PMCID: PMC11605106.
5. Velamala BS, Gautam S, Lohani S. Role of Artificial Intelligence in Reducing Error Rates in Radiology: A Scoping Review. *Cureus*. 2025 Sep 10;17(9):e91957. doi: 10.7759/cureus.91957. PMID: 41080241; PMCID: PMC12512053.
6. Singh Y, Hathaway QA, Keishing V, Salehi S, Wei Y, Horvat N, Vera-Garcia DV, Choudhary A, Mula Kh A, Quaia E, Andersen JB. Beyond Post hoc Explanations: A Comprehensive Framework for Accountable AI in Medical Imaging Through Transparency, Interpretability, and Explainability. *Bioengineering (Basel)*. 2025 Aug 15;12(8):879. doi: 10.3390/bioengineering12080879. PMID: 40868392; PMCID: PMC12383817.