



The Role of Explainable AI and Ethical Considerations in Clinical Radiology: A Mini-Review

Noah John*

School of Clinical Sciences, Western Biomedical College, Canada

***Corresponding author:** Noah John, School of Clinical Sciences, Western Biomedical College, Canada

Mini Review

Volume 10 Issue 1

Received Date: March 04, 2026

Published Date: March 24, 2026

Abstract

Artificial intelligence (AI) has rapidly become integrated into clinical radiology, supporting image acquisition, lesion detection, disease classification, workflow optimization, and decision support. While deep learning algorithms frequently achieve diagnostic performance comparable to experienced radiologists for selected imaging tasks, many of these models operate as "black boxes," limiting clinician confidence and hindering widespread clinical acceptance. Explainable artificial intelligence (XAI) has emerged as a promising solution by providing transparent insights into algorithmic decision-making through visualization techniques, feature attribution methods, and interpretable prediction models. Such approaches can improve physician trust, facilitate error identification, and enhance collaboration between AI systems and radiologists. Alongside technological advances, ethical challenges remain central to responsible AI implementation. Concerns regarding algorithmic bias, patient privacy, informed consent, accountability, regulatory oversight, and fairness require careful consideration before routine clinical deployment. Existing evidence suggests that explainability alone does not guarantee trustworthy AI; instead, technical transparency must be combined with rigorous validation, diverse datasets, continuous monitoring, and ethical governance. Recent international recommendations increasingly emphasize human oversight, standardized reporting, and multidisciplinary collaboration during AI development and evaluation. This mini-review summarizes current developments in explainable AI for clinical radiology while highlighting major ethical principles and implementation challenges. It also discusses emerging research directions that may facilitate safer, more transparent, and clinically meaningful integration of AI technologies into modern radiological practice.

Keywords: Explainable artificial intelligence; Clinical radiology; Medical imaging; Ethics; Machine learning; Diagnostic decision support

Abbreviations: AI: Artificial Intelligence; XAI: Explainable Artificial Intelligence; CT: Computed Tomography; MRI: Magnetic Resonance Imaging; Grad-CAM: Gradient-Weighted Class Activation Mapping; LIME: Local Interpretable Model-Agnostic Explanations; SHAP: Shapley Additive Explanations.

Introduction

Artificial intelligence has transformed many aspects of diagnostic imaging by enabling automated interpretation of radiographs, computed tomography (CT), magnetic resonance imaging (MRI), ultrasound,

and digital pathology images. Deep convolutional neural networks have demonstrated remarkable performance in image classification and lesion detection across multiple clinical applications. Despite these advances, many high-performing models provide limited information regarding the reasoning behind their predictions. This lack of transparency creates challenges for clinical acceptance, particularly when imaging findings influence high-stakes medical decisions. Explainable AI seeks to bridge this gap by generating interpretable outputs that allow clinicians to understand, verify, and critically evaluate algorithmic recommendations while maintaining patient safety and professional accountability [1].

Explainable AI in Radiological Decision Support

Explainable AI encompasses computational methods that clarify how machine learning models arrive at specific predictions. Common techniques include saliency maps, Gradient-weighted Class Activation Mapping (Grad-CAM), Local Interpretable Model-Agnostic Explanations (LIME), and SHapley Additive exPlanations (SHAP). These methods identify image regions or features that contribute most strongly to diagnostic classifications. In radiology, explainability supports quality assurance by allowing radiologists to verify whether algorithms focus on clinically relevant anatomical structures rather than unrelated image artifacts. Such transparency is particularly valuable in breast imaging, chest radiography, neuroimaging, and oncologic imaging, where subtle imaging characteristics influence diagnostic confidence. Explainable outputs may also facilitate educational training by illustrating imaging features associated with specific disease patterns [2].

Ethical Challenges in Clinical AI Implementation

Ethical concerns extend beyond model accuracy. Algorithmic bias remains a significant challenge because training datasets may underrepresent certain demographic groups, imaging equipment, or disease prevalence patterns. As a result, diagnostic performance may vary across different patient populations, potentially contributing to healthcare disparities. Patient privacy represents another critical issue since AI development often requires access to large imaging datasets linked with clinical information. Robust de-identification procedures, secure data governance, and compliance with regional privacy regulations are essential. Questions surrounding legal responsibility also remain unresolved. When AI-assisted diagnoses contribute to medical errors, determining accountability among software developers, healthcare institutions, and clinicians requires clearly defined regulatory frameworks [3].

Current Advances and Future Perspectives

Recent research increasingly combines explainability with uncertainty estimation, federated learning, and continuous model monitoring. Federated learning enables collaborative model development across institutions while minimizing direct patient data sharing, thereby strengthening privacy protection. Simultaneously, uncertainty-aware AI systems communicate prediction confidence, enabling radiologists to identify cases requiring additional clinical review. Professional organizations are also promoting standardized reporting guidelines for AI validation, emphasizing external validation, reproducibility, fairness assessment, and post-deployment surveillance. Rather than replacing radiologists, explainable AI is increasingly viewed as an intelligent decision-support tool that complements human expertise through transparent and evidence-based collaboration [4].

Limitations and Research Gaps

Although explainable AI offers promising benefits, several limitations remain. Many explanation techniques provide approximate rather than definitive interpretations of model behavior and may themselves introduce bias or misleading visualizations. There is limited consensus regarding standardized evaluation metrics for explanation quality, making comparisons across studies difficult. Most published AI models continue to rely on retrospective datasets from single institutions, limiting generalizability across diverse healthcare settings. Prospective multicenter validation, standardized benchmarking protocols, and regulatory guidance are needed before explainable AI can achieve widespread routine clinical implementation. Future investigations should also evaluate whether improved explainability directly enhances diagnostic accuracy, physician confidence, patient outcomes, and healthcare efficiency [5].

Conclusion

Explainable AI represents an important advancement toward trustworthy and clinically acceptable artificial intelligence in radiology. By increasing transparency and supporting clinician interpretation, explainable models have the potential to strengthen diagnostic confidence and facilitate safer AI-assisted healthcare. Nevertheless, technical innovation must be accompanied by ethical governance addressing fairness, privacy, accountability, and regulatory compliance. Continued collaboration among radiologists, computer scientists, ethicists, healthcare administrators, and policymakers will be essential to establish reliable, transparent, and patient-centered AI systems capable of supporting future radiological practice.

References

1. McKinney SM, Sieniek M, Ledsam J R, Godbole V, Mostofi H, et al. (2020) International evaluation of an AI system for breast cancer screening. *Nature* 577: 89-94.
2. Char DS, Shah NH, Magnus D (2018) Implementing machine learning in healthcare-Addressing ethical challenges. *N Engl J Med* 378: 981-983.
3. Geras KJ, Mann RM, Moy L (2019) Artificial intelligence for mammography and digital breast tomosynthesis: Current concepts and future perspectives. *Radiology* 293(2): 246-259.
4. Char DS, Shah NH, Magnus D (2018) Implementing machine learning in health care—Addressing ethical challenges 378(11): 981-983.
5. Erickson BJ, Korfiatis P, Akkus Z, Kline TL (2017) Machine learning for medical imaging. *Radiographics* 37(2): 505-515.