Advanced Mechanical and Mechatronic Systems

Review
Volume 1(1), Year, 52–66
https://doi.org/10.53623/amms.v1i1.692

Explainable Artificial Intelligence (XAI) in Medical Imaging: Techniques, Applications, Challenges, and Future Directions

Purwono Purwono*, Annastasya Nabila Elsa Wulandari, Khoirun Nisa

Department of Informatics, Universitas Harapan Bangsa, Purwokerto, Indonesia

*Correspondence: purwono@uhb.ac.id

SUBMITTED: 19 May 2025; REVISED: 13 June 2025; ACCEPTED: 15 June 2025

ABSTRACT: The integration of Explainable Artificial Intelligence (XAI) into medical imaging is pivotal in addressing the "black-box" limitations of deep learning models, which often hinder clinical trust and regulatory approval. This review provides a comprehensive examination of XAI techniques that enhance interpretability and transparency in diagnostic imaging applications. Key approaches such as feature visualization (Grad-CAM, Integrated Gradients), attention mechanisms, symbolic reasoning, and example-based methods—are explored alongside their practical implementations. Specific cases in cardiac imaging, cancer diagnostics, and brain lesion segmentation illustrate the value of XAI in improving clinical decision-making and patient care. Moreover, the review highlights major challenges, including the trade-off between accuracy and interpretability, ethical and legal constraints, integration barriers within clinical workflows, and the complexity of medical data. To address these issues, future research directions are proposed, including the development of more robust example-based models, ethical frameworks, generalizable architectures, advanced visualization techniques, and interdisciplinary collaboration. With continued refinement and responsible deployment, XAI systems can enable AI models to become not only accurate but also interpretable and clinically relevant. This paper underscores the transformative potential of XAI in building trustworthy, transparent, and effective AI-driven diagnostic tools aligned with the practical demands of modern healthcare systems.

KEYWORDS: Explainable Artificial Intelligence (XAI); medical imaging; interpretability; clinical trust; deep learning

1. Introduction

Artificial Intelligence (AI) became a transformative force in medical imaging, offering substantial improvements in image-based diagnosis, disease detection, and clinical workflow optimization [1, 2]. With the rapid advancement of deep learning (DL) models, significant breakthroughs were achieved in tasks such as image classification, segmentation, and anomaly detection. These AI-powered systems accelerated diagnostic processes, enhanced accuracy, and supported decision-making in complex clinical scenarios [3].

Despite these advancements, a major barrier continued to hinder the widespread adoption of AI in clinical practice: the lack of transparency and interpretability of high-performing models [4]. Many DL systems operated as "black boxes," providing predictions without revealing the underlying decision-making process. This lack of explainability was particularly concerning in healthcare, where understanding the rationale behind diagnostic outcomes remained critical for both clinicians and patients.

The inability to explain AI-generated decisions undermined clinical trust and introduced ethical, legal, and safety concerns, especially in high-stakes domains such as radiology, oncology, and neurology [5, 6]. Healthcare professionals needed to validate automated recommendations, ensure alignment with medical guidelines, and communicate findings effectively to patients. As a result, there was growing demand for AI systems that were not only accurate but also transparent and interpretable.

XAI emerged to address these concerns by providing interpretable insights into model behavior. XAI techniques, including feature visualization, attention mechanisms, rule-based reasoning, and post hoc explanation methods, aimed to identify the features or image regions most responsible for a model's prediction [7]. By improving interpretability, XAI facilitated clinical acceptance, supported diagnostic decision-making, and helped bridge the gap between opaque AI systems and human reasoning.

This article presented a systematic and comprehensive review of XAI methods in medical imaging by evaluating widely used techniques and their applications in various clinical fields such as cardiology, oncology, neurology, and dermatology. A structured classification was developed to associate each XAI approach with specific clinical tasks and imaging modalities. Comparative analyses were conducted on interpretability, performance trade-offs, and the ethical and regulatory aspects that accompanied the use of these methods. Some approaches that remained rarely implemented in clinical practice, such as example-based methods and symbolic reasoning, were highlighted, with emphasis placed on the importance of integrating regulatory considerations and practical implementation. The article concluded by mapping future research directions focused on developing reliable, generalizable, and ethically sound XAI systems for effective integration into clinical practice.

2. Fundamentals of Explainable Artificial Intelligence

XAI was an emerging field focused on enhancing the transparency and interpretability of AI models. Its primary goal was to enable users, particularly in high-stakes fields such as healthcare, to understand, trust, and effectively manage the decisions produced by AI systems. Unlike traditional "black box" models that offered limited insight into their internal logic, XAI methods aimed to reveal and communicate the reasoning behind model outputs in ways that were accessible to human users [8].

A foundational principle of XAI involved the distinction between transparency and interpretability. Transparency referred to how clearly the internal mechanisms of a model could be examined, while interpretability related to how easily humans could understand the relationships between inputs and outputs in a model's decision-making process [9, 10]. In medical imaging, both attributes were essential because clinical decisions required clear justification and alignment with established diagnostic protocols.

XAI included a variety of explanation types to address different user needs and technical contexts. For instance, techniques such as Local Interpretable Model-Agnostic Explanations

(LIME) and SHapley Additive exPlanations (SHAP) were model-agnostic methods applied after training to interpret individual predictions [11]. Alternatively, some models were inherently interpretable by design, such as decision trees and rule-based systems, which provided native transparency and were often preferred in sensitive or regulated domains [12].

Another important aspect of XAI was its sensitivity to context. The effectiveness of an explanation depended on its relevance to the target audience and the specific application domain. Technical explanations might have been suitable for data scientists but too complex for clinicians or patients. Therefore, successful implementation of XAI required careful consideration of who the end users were and what level of understanding was necessary [13]. This ensured that explanations were not only technically accurate but also meaningful in practice.

In medical imaging, the application of XAI methods required particular attention to the diversity of imaging modalities, each with unique visual characteristics and clinical implications. To provide a visual overview of this diversity, Figure 1 presented illustrations of some of the most common types of medical imaging that formed the basis for the development and application of XAI approaches.

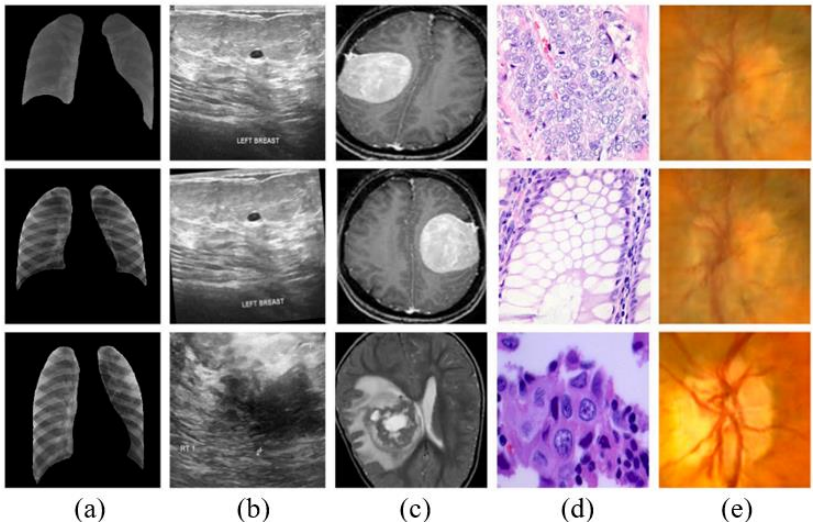

Figure 1. Representative medical imaging modalities: (a) chest X-ray, (b) breast ultrasonography, (c) brain MRI, (d) histopathological tissue imaging, and (e) retinal fundus. These modalities demonstrate the diversity of visual data that needs to be interpreted by the XAI method in support of clinical decisions [14].

3. XAI Techniques in Medical Imaging

The classification of XAI techniques in this study was based on the main interpretability strategies used to explain model prediction results. These approaches included feature visualization, attention mechanisms, rule-based reasoning, example-based methods, and post hoc explanations. This taxonomy was developed by referencing findings from recent scientific literature and was contextually adapted to meet interpretability needs in clinical practice, particularly in the field of medical imaging, which required transparency, accountability, and clarity in decision-making processes.

3.1. Feature visualization techniques.

Feature visualization techniques were essential for enhancing the interpretability of deep learning models in medical imaging. These methods provided visual insights into the regions

or features of an input image that contributed most to a model's decision, thereby improving transparency and clinical trust. Grad-CAM was one of the most widely used techniques, generating class-specific heatmaps that highlighted image areas with the greatest influence on the model's prediction [15]. It was applied in various clinical contexts, such as detecting attention-deficit/hyperactivity disorder using EEG data [15], identifying cataracts from fundus images [16], and diagnosing hip fractures in pelvic radiographs [17].

Other methods, such as Integrated Gradients and SmoothGrad, increased the precision and reliability of visual explanations. Integrated Gradients attributed model predictions to input features by integrating gradients along a path from a baseline to the actual input, enabling quantification of feature importance [18, 19]. SmoothGrad enhanced saliency maps by adding noise to inputs and averaging the gradients, thereby reducing visual artifacts and clarifying relevant feature contributions [18].

Additional approaches included activation maps and deconvolution methods, which visualized internal feature activations within convolutional neural networks. These helped illustrate how models processed information across layers. Layered Grad-CAM extended traditional Grad-CAM by aggregating visual explanations from multiple layers, improving interpretability in complex architectures such as Feature Pyramid Networks, which were used in polyp segmentation [20]. Example-based techniques also supported explainability by retrieving and displaying similar cases from training data, helping clinicians understand and validate AI predictions in practice [21].

3.2. Attention mechanisms.

Attention mechanisms were an important component of XAI techniques in medical imaging, offering improved interpretability by allowing models to focus on the most relevant regions of input data. These mechanisms were generally categorized as post hoc attention, which analyzed pre-trained networks to uncover decision logic, and trainable attention, which guided model learning toward informative regions during training [22]. This categorization enabled attention-based models to enhance both accuracy and transparency in diagnostic applications.

In classification and segmentation tasks, attention mechanisms demonstrated significant utility. Models that incorporated attention achieved high accuracy in diagnosing diseases such as Covid-19, breast cancer, lung cancer, and retinal conditions [23]. For example, Covid-19 radiography classification achieved up to 98 percent accuracy, while lung cancer classification reached up to 99.8 percent. In segmentation, architectures such as MDSU-Net employed dual attention and attention gates to enhance feature representation and delineation quality [24]. Furthermore, self-attention mechanisms in transformer-based models identified critical regions in medical images, thereby increasing interpretability and model trustworthiness [25].

Several explainability techniques were used in conjunction with attention-based models. Grad-CAM was frequently employed to visualize relevant image regions associated with predictions [23, 26], while LIME and SHAP provided model-agnostic interpretations by attributing predictions to input features. More recently, techniques such as Attention-Gradient Class Activation Mapping (A-GCAM) were introduced to analyze attention attribution in a more transparent way [27]. Despite these advancements, challenges remained, including the complexity of attention modules and the lack of standardized metrics for clinical validation [28]. Ongoing research was necessary to develop attention mechanisms that were both accurate and clinically interpretable.

3.3. Rule-based and symbolic XAI approaches.

Rule-based and symbolic approaches played a critical role in enhancing the transparency, reliability, and clinical relevance of AI systems in medical imaging. These techniques were well suited for healthcare applications due to their inherent interpretability and alignment with clinical guidelines. Rule-based systems provided explicit, understandable decision logic, allowing clinicians to verify outcomes and validate model behavior against established protocols.

A notable example was the neurosymbolic system developed for vertebral compression fracture detection in computed tomography scans. This system combined deep learning for vertebral segmentation with a shape-based algorithm that analyzed vertebral height distributions to define diagnostic rules. It achieved 96 percent accuracy and 91 percent sensitivity, showing that rule-based methods could perform comparably to black-box models while offering greater interpretability [29]. Another example was the Adaptive Neuro-Fuzzy Inference System (ANFIS), used in Intensity-Modulated Radiation Therapy planning. ANFIS combined neural networks with rule-based reasoning to optimize treatment plan selection and provide interpretable justifications for decisions, thereby increasing clinical trust [30].

Symbolic XAI techniques extended interpretability further by employing logical or rule-based representations. For instance, one deep learning framework for cancer image classification incorporated symbolic reasoning to deliver high accuracy (97.72 percent) along with user-adaptive explanations for healthcare professionals [31]. Another model for contrast phase detection in abdominal CT scans used a hybrid approach, combining deep learning with rule-based methods. It achieved 92.3 percent accuracy and applied Shapley values to explain the relevance of radiodensity features, improving transparency and clinical usefulness [32].

3.4. Example-based and case-based reasoning.

Example-based and case-based reasoning offered intuitive and clinically relevant explanations by referencing specific past cases. These approaches helped users understand AI model decisions by illustrating how similar inputs had been classified. Prototype-based methods identified representative examples during training and compared new inputs to these prototypes. Models such as ProtoPNet were successfully applied in medical fields like brain tumor classification, offering interpretability without compromising accuracy [33].

Retrieval-based techniques also provided explanations by retrieving similar examples from a reference dataset. This enabled clinicians to validate AI predictions by comparing them to previously seen and verified cases [34]. Another powerful method was counterfactual explanation, which showed how small changes to input data could alter predictions. This highlighted model decision boundaries and offered insights into alternative diagnostic outcomes [34]. Case-Based Reasoning (CBR) extended this approach by using analogical reasoning to address new clinical problems based on past cases. CBR systems provided visual explanations by displaying similar cases and identifying shared characteristics. In breast cancer diagnosis, for example, CBR systems helped clinicians classify findings by referencing comparable historical examples [35]. Hybrid models that combined CBR with deep learning demonstrated improved accuracy and interpretability, as seen in mammogram analysis systems [36].

3.5. Surrogate and post-hoc explainers.

Surrogate and post hoc explainers were vital to XAI in medical imaging. As AI-based diagnostic tools became more widespread, ensuring transparency and interpretability was essential for clinical adoption and trust. These XAI techniques made the rationale behind model predictions accessible, allowing clinicians to understand and justify automated outputs [37]. Post hoc explainers were applied after training and did not interfere with model performance. One such method was counterfactual explanation, which altered input data to observe changes in predictions. While informative, this technique was computationally demanding, especially in high-dimensional medical imaging contexts [38]. Saliency mapping was another common approach, highlighting influential image regions. Techniques such as integrated gradients and guided backpropagation were widely used, each with strengths in localization and computational efficiency [39].

In contrast, surrogate explainers aimed to approximate complex models using simpler, interpretable alternatives. LIME was a well-known example, constructing locally linear models to explain specific predictions [40]. Similarly, rule-based surrogates mimicked the decision logic of complex models through predefined rules, making them especially valuable in regulated medical environments [41]. These surrogate models played a key role in improving clinician trust and supporting compliance with ethical and regulatory standards. Table 1 shows Comparative summary of explainable AI (XAI) techniques in medical imaging.

Table 1. Comparative summary of explainable AI (XAI) techniques in medical imaging.

Engineering	Medical Domain	Explanation Type	Strengths	Limitations
Grad-CAM (Gradient-weighted Class Activation Mapping) [42].	Radiology (MRI, CT, X-ray, Histology)	Gradient-based Heatmap (visual class localisation)	Highlights important areas in the image, compatible with various CNN architectures	Depends on the last layer of the CNN, lacks precision for complex pixel-level segmentation
SHAP [43].	Oncology (Breast, Lung, Colon)	Shapley value (global and local feature attribution)	Computationally expensive, interpretation of visualisations can be complex for non-technical people	Provides quantitative contribution of each pixel to prediction, accurate, consistent in explanation
LIME [44].	Radiology	Local Surrogate (superpixel and local interpretable model)	Flexible, applicable to all models, visually clear	Parameter sensitive, results may vary between executions
Attention U-Net [42].	Radiology (Abdomen, Brain, Chest)	Intrinsic explanation	Focus on important anatomical areas, improves segmentation	Architecture must be modified, can overfit on small data
CBR (Case Retrieval & Matching) [45].	Paediatric Oncology (Kidney Tumour)	Analogue reasoning (similarity-based retrieval)	Selects most similar cases from case base for training, improves generalisation	Depends on quality of case base, performance may degrade when case base is limited
ANFIS (Fuzzy Rule-Based + Neural Network) [46].	Neurology (Brain Tumour)	Fuzzy Rule-Based (interpretable)	Fuzzy rule-based clear interpretation, suitable for expert knowledge integration	Less efficient for big data, needs manual tuning of rules and parameters
Inception v3 + Integrated Gradients [47].	Radiology (Pneumonia)	Gradient-based Attribution (pixel-level)	Provides visualisation of important areas in X-ray images, improves clinical confidence	Needs baseline image and additional computation, sensitivity to baseline selection

4. Applications of XAI in Medical Imaging

4.1. Cardiac imaging.

The application of AI in cardiac imaging significantly advanced diagnostic capabilities by offering improved accuracy, automated quantification, and enhanced workflow efficiency. AI algorithms were widely implemented in coronary computed tomography angiography (CCTA) to evaluate calcium scores, quantify coronary stenosis, and analyze plaque composition [48]. Similarly, in cardiac magnetic resonance imaging (CMR), AI supported the segmentation of cardiac chambers and the characterization of myocardial tissue, which were essential for diagnosing myocardial infarction and cardiomyopathy [49]. In echocardiography, AI facilitated the segmentation of cardiac structures to assess valvular function and wall motion abnormalities [48]. Despite these advancements, the lack of interpretability in conventional AI systems limited their integration into clinical workflows, where explainability was essential for clinical acceptance and regulatory compliance.

XAI addressed this limitation by providing transparency into the model's decision-making process, enabling clinicians to understand, validate, and trust AI-generated outputs. In diagnostic applications, XAI techniques such as saliency maps, attention mechanisms, and post hoc attribution methods helped identify the specific regions or features that contributed to model predictions. These explanations were particularly important in high-stakes scenarios such as identifying myocardial ischemia or classifying coronary lesions. Additionally, AI demonstrated substantial benefits in improving workflow efficiency by reducing image acquisition and post-processing time, which accelerated diagnostic reporting without compromising accuracy [50]. XAI also contributed to reducing interobserver variability by offering consistent, interpretable outputs that aligned with clinical reasoning [50]. Therefore, the integration of XAI in cardiac imaging represented a critical step toward safe, efficient, and trustworthy AI-assisted diagnostics.

4.2. Cancer diagnosis.

XAI became an essential component in cancer diagnostics by offering interpretable insights that enhanced the clinical trustworthiness of AI systems. Across various cancer types, XAI was integrated into both image-based and biomarker-based diagnostic workflows to address the limitations of black-box models. In oral cancer detection, the EXAIOC framework combined convolutional neural networks (CNNs) with fuzzy logic to manage data uncertainties. Techniques such as Layer-wise Relevance Propagation (LRP) and Grad-CAM generated visual explanations that improved interpretability and supported clinical decision-making [51]. In breast cancer detection, explainable deep recurrent convolutional neural networks (RCNNs) enhanced with transfer learning effectively captured both spatial and temporal features in mammographic images [52].

XAI also demonstrated value in prostate, renal, and liver cancer diagnostics through interpretable modeling techniques. In prostate cancer, LIME and shape analysis were used to interpret MRI-based predictions, thereby addressing the opacity of deep learning models and increasing clinical trust [53]. For renal cancer detection, CNNs applied to high-resolution medical images, combined with Grad-CAM and LIME, achieved high diagnostic accuracy while providing visual explanations to help clinicians understand model outputs [54]. Similarly, the integration of U-Net with LIME in liver cancer detection enabled more accurate

segmentation and clearer model interpretation, thereby supporting physician decision-making [55].

5. Challenges in XAI for Medical Imaging

The implementation of XAI in medical imaging offered great potential to enhance diagnostic precision and build clinical trust. Nevertheless, several critical challenges needed to be addressed to ensure that XAI systems were not only accurate but also ethical, transparent, and practical for clinical use. As AI technologies became more advanced, the complexity of the models increased, making it more difficult to design systems that were both interpretable and suitable for deployment in healthcare settings. A major obstacle to XAI adoption was the opaque nature of many high-performing AI models, especially those based on deep learning. These models often generated predictions without offering insight into how or why decisions were made [56, 57]. In medical environments, where accountability and justification were essential, this lack of transparency could diminish trust among clinicians and patients [58]. Without clear interpretability, AI tools might be viewed as unreliable or unsuitable for clinical decision-making.

Another central challenge involved managing the trade-off between interpretability and predictive accuracy. Complex models that delivered high performance were often difficult to interpret, whereas simpler models were easier to explain but less accurate [59, 60]. This compromise created a tension between achieving state-of-the-art performance and ensuring clinical usability. Striking the right balance remained a core focus of XAI research, particularly in critical diagnostic domains. Incorporating XAI into existing clinical workflows also presented significant challenges. To be useful, AI-generated explanations needed to be timely, clinically relevant, and understandable to medical professionals working under time pressure [61]. If explanations were too abstract, delayed, or disconnected from clinical logic, they were unlikely to be used. Successful integration required collaboration among AI developers, clinicians, and system engineers to ensure that outputs were practical and aligned with healthcare needs.

Ethical and regulatory issues further complicated the use of XAI in medical imaging. Concerns such as algorithmic bias, patient data privacy, and the risk of incorrect diagnoses had to be addressed with caution [62]. These challenges were intensified by the fact that ethical guidelines and legal regulations had not kept pace with the rapid evolution of AI technologies. Without proper oversight, even explainable models might cause harm, particularly if their outputs were based on biased or low-quality data. Lastly, fostering trust across stakeholders, including clinicians, patients, healthcare administrators, and regulators, was critical for the success of XAI in medical applications [63]. Clear and open communication about how models were developed, trained, and validated was essential. Furthermore, technical and data-related limitations had to be considered. Medical data was often complex, high-dimensional, and variable. Designing interpretable models that could effectively learn from such data remained a continuing challenge. Ensuring data integrity, governance, and contextual relevance was key to building dependable and explainable AI systems for clinical use [64].

6. Future Directions

Future research in XAI for medical imaging is expected to focus on refining example-based techniques, which remained underutilized in clinical practice. These methods offered potential

for improving interpretability by presenting concrete, relatable examples that could justify AI model predictions. Moving forward, researchers needed to develop more robust and scalable example-based models that generated intuitive explanations aligned with clinical reasoning [65]. Emphasis should be placed on ensuring these methods were both accurate and accessible to healthcare professionals in real-world settings. Addressing ethical and technical challenges would be critical to the responsible advancement of XAI in medical imaging. Future research should prioritize the development of ethically sound models that aligned with the values and needs of the healthcare sector [66]. This included addressing persistent concerns related to algorithmic bias, data privacy, and model security [67]. Ensuring that XAI methods complied with emerging ethical frameworks would be essential for earning the trust of both clinicians and patients.

Improving the generalizability and diagnostic performance of explainable models remained a central research objective. Combining powerful architectures such as ResNet50 with XAI techniques showed potential in improving transparency and performance simultaneously [68]. Future investigations should build upon these efforts by exploring hybrid models that balanced interpretability with high predictive accuracy, ensuring they were adaptable across various imaging modalities and patient populations. The convergence of XAI with Big Data Analytics (BDA) and the Internet of Medical Things (IoMT) represented a transformative direction in healthcare innovation. Leveraging these technologies could facilitate the development of scalable XAI systems capable of processing large, heterogeneous datasets for real-time decision support [69]. Research should focus on optimizing XAI algorithms to efficiently manage and interpret high-volume data streams, thereby supporting personalized and cost-effective care at scale. There was also an urgent need for standardized evaluation metrics and regulatory frameworks to guide the responsible deployment of XAI systems in clinical environments [64]. Future work should aim to establish consensus on best practices for model validation, interpretability assessment, and ethical compliance. Regulatory bodies and research institutions must collaborate to create guidelines that ensured fairness, transparency, and accountability in AI-driven diagnostics.

Promoting multidisciplinary collaboration was essential for the successful integration of XAI into clinical practice. Partnerships among AI developers, radiologists, and other healthcare professionals could ensure that XAI models were designed with clinical utility in mind [70]. Future initiatives should support co-development approaches that aligned technical innovation with the workflows and decision-making processes of end users, thereby improving model relevance and adoption. Finally, the development of advanced visualization techniques held great promise for enhancing the interpretability of AI models. Pixel-level heatmaps and layered visual explanations could help clinicians better understand model predictions and their underlying rationale [71]. Future research should focus on improving the resolution, clarity, and interactivity of these visual outputs to deliver more actionable insights that supported clinical decision-making.

7. Conclusions

XAI was an essential enabler for the responsible and effective adoption of AI in medical imaging. Through a detailed review of current techniques and applications, this article demonstrated how XAI could bridge the gap between high-performing deep learning models and the need for interpretability in clinical environments. Despite notable progress, significant

challenges remained, including data limitations, ethical concerns, and model complexity. Future research must prioritize the development of interpretable, generalizable, and ethically sound XAI models that were compatible with clinical workflows. Furthermore, advancements in visualization tools, example-based reasoning, and collaborative model design would be critical for improving clinical usability. Ultimately, the continued evolution of XAI in medical imaging had the potential to foster greater trust, transparency, and accountability in AI-assisted healthcare delivery.

Acknowledgments

The authors expressed their sincere appreciation to colleagues and clinical experts who provided valuable insights that enriched the perspectives presented in this review. They also acknowledged the institutional support from Universitas Harapan Bangsa, which enabled access to relevant literature and resources. The authors were grateful to the anonymous reviewers for their constructive comments and suggestions, which helped improve the quality and clarity of the manuscript.

Author Contribution

All authors contributed to the development and completion of this study. Purwono and Annastasya Nabila Elsa Wulandari were responsible for the conceptualization of the research. Purwono and Khoirun Nisa designed the methodology and conducted the data collection and analysis. The manuscript was written collaboratively by Purwono, Annastasya Nabila Elsa Wulandari, and Khoirun Nisa. Each author reviewed and approved the final version of the manuscript.

Competing Interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

References

[1] Muhammad, D.; Bendechache, M. (2024). Unveiling the black box: A systematic review of Explainable Artificial Intelligence in medical image analysis. *Computational and Structural Biotechnology Journal*, *24*, 542–560. https://doi.org/10.1016/j.csbj.2024.08.005.

[2] Narayankar, P.; Baligar, V.P. (2024). Explainability of Brain Tumor Classification Based on Region. 2024 International Conference on Emerging Technologies in Computer Science for Interdisciplinary Applications, 1–6. https://doi.org/10.1109/ICETCS61022.2024.10544289.

[3] Patrício, C.; Neves, J.C.; Teixeira, L.F. (2024). Explainable Deep Learning Methods in Medical Image Classification: A Survey. *ACM Computing Surveys*, *56*(4), 1–41. https://doi.org/10.1145/3625287.

[4] Hardani, D.N.K.; Ardiyanto, I.; Adi Nugroho, H. (2024). Decoding brain tumor insights: Evaluating CAM variants with 3D U-Net for segmentation. *Communications in Science and Technology*, *9*(2), 262–273. https://doi.org/10.21924/cst.9.2.2024.1477.

[5] Salahuddin, Z.; Woodruff, H.C.; Chatterjee, A.; Lambin, P. (2022). Transparency of deep neural networks for medical image analysis: A review of interpretability methods. *Computers in Biology and Medicine*, *140*, 105111. https://doi.org/10.1016/j.compbiomed.2021.105111.

[6] Hachi, S.; Sabri, A. (2024). Development of an Explainable AI (XAI) system for the interpretation of ViT in the diagnosis of medical images. 2024 3rd International Conference on Embedded

Systems and Artificial Intelligence (ESAI). 1–8. https://doi.org/10.1109/ESAI62891.2024.10913833.

[7] Inigo, B.; et al. (2025). An intrinsically explainable approach to detecting vertebral compression fractures in CT scans via neurosymbolic modeling. In Medical Imaging 2025: Image Processing; Colliot, O.; Mitra, J., Eds.; SPIE: Bellingham, WA, USA, p. 101. https://doi.org/10.1117/12.3047426.

[8] DeSimone, H.; Leon-Espinosa, M. (2024). Explainable AI: The Quest for Transparency in Business and Beyond. 7th International Conference on Information and Computer Technologies, 532–538. https://doi.org/10.1109/ICICT62343.2024.00093.

[9] Keppeler, V.; Lederer, M.; Leucht, U.A. (2022). Explainable Artificial Intelligence. In Encyclopedia of Data Science and Machine Learning; IGI Global: Hershey, PA, USA, pp. 1667–1684. https://doi.org/10.4018/978-1-7998-9220-5.ch100.

[10] Dhar, A.; Gupta, S.; Kumar, E.S. (2024). A Comprehensive Review of Explainable AI Applications in Healthcare. 15th International Conference on Computing Communication and Networking Technologies, 1–8. https://doi.org/10.1109/ICCCNT61001.2024.10725578.

[11] Bhushan, S.; Dixit, S. (2024). Explainable AI for Shaping Adoption of Artificial Intelligence. *Asian Conference on Intelligent Technologies*, 1–6. https://doi.org/10.1109/ACOIT62457.2024.10939416.

[12] Khakurel, U.B.; Rawat, D.B. (2022). Evaluating explainable artificial intelligence (XAI): algorithmic explanations for transparency and trustworthiness of ML algorithms and AI systems. In Artificial Intelligence and Machine Learning for Multi-Domain Operations Applications IV; Pham, T.; Solomon, L.; Hohil, M.E., Eds.; SPIE: Bellingham, WA, USA, p. 7. https://doi.org/10.1117/12.2620598.

[13] Nyrup, R.; Robinson, D. (2022). Explanatory pragmatism: a context-sensitive framework for explainable medical AI. *Ethics and Information Technology*, *24*(1), 13. https://doi.org/10.1007/s10676-022-09632-3.

[14] Ullah, N.; Guzmán-Aroca, F.; Martínez-Álvarez, F.; De Falco, I.; Sannino, G. (2025). A novel explainable AI framework for medical image classification integrating statistical, visual, and rule-based methods. *Medical Image Analysis*, *105*, 103665. https://doi.org/10.1016/j.media.2025.103665.

[15] Latifi, B.; Amini, A.; Motie Nasrabadi, A. (2024). Siamese based deep neural network for ADHD detection using EEG signal. *Computers in Biology and Medicine*, *182*, 109092. https://doi.org/10.1016/j.compbiomed.2024.109092.

[16] Shah, H.; Patel, R.; Hegde, S.; Dalvi, H. (2023). XAI Meets Ophthalmology: An Explainable Approach to Cataract Detection Using VGG-19 and Grad-CAM. IEEE Pune Section International Conference (PuneCon), 1–8. https://doi.org/10.1109/PuneCon58714.2023.10450053.

[17] Chung, S.-L.; Cheng, C.-T.; Liao, C.-H.; Chung, I.-F. (2025). Patch-based feature mapping with generative adversarial networks for auxiliary hip fracture detection. *Computers in Biology and Medicine*, *186*, 109627. https://doi.org/10.1016/j.compbiomed.2024.109627.

[18] Papanastasopoulos, Z.; et al. (2020). Explainable AI for medical imaging: deep-learning CNN ensemble for classification of estrogen receptor status from breast MRI. In Medical Imaging 2020: Computer-Aided Diagnosis; Hahn, H.K.; Mazurowski, M.A., Eds.; SPIE: Bellingham, WA, USA, p. 52. https://doi.org/10.1117/12.2549298.

[19] Narayankar, P.; Baligar, V.P. (2024). Explainability of Brain Tumor Classification Based on Region. *International Conference on Emerging Technologies in Computer Science for Interdisciplinary Applications*, 1–6. https://doi.org/10.1109/ICETCS61022.2024.10544289.

[20] Javali, S.M.; Surya Upadhyayula, R.; De, T. (2022). Comparative study of xAI layer-wise algorithms with a Robust Recommendation framework of Inductive Clustering for Polyp

Segmentation and Classification. International Seminar on Machine Learning, Optimization, and Data Science (ISMODE), 325–330. https://doi.org/10.1109/ISMODE53584.2022.9743003.

[21] Fontes, M.; De Almeida, J.D.S.; Cunha, A. (2024). Application of Example-Based Explainable Artificial Intelligence (XAI) for Analysis and Interpretation of Medical Imaging: A Systematic Review. *IEEE Access*, *12*, 26419–26427. https://doi.org/10.1109/ACCESS.2024.3367606.

[22] Shin, H.; Lee, J.; Eo, T.; Jun, Y.; Kim, S.; Hwang, D. (2020). The Latest Trends in Attention Mechanisms and Their Application in Medical Imaging. *Journal of the Korean Society of Radiology*, *81*(6), 1305. https://doi.org/10.3348/jksr.2020.0150.

[23] Muoka, G.W.; Yi, D.; Ukwuoma, C.C.; Martin, M.D.; Aydin, A.A.; Al-Antari, M.A. (2023). A Novel Attention-based Explainable Deep Learning Framework Towards Medical Image Classification. *7th International Symposium on Innovative Approaches in Smart Technologies*, 1–8. https://doi.org/10.1109/ISAS60782.2023.10391289.

[24] Zhou, Y.; Kang, X.; Ren, F.; Lu, H.; Nakagawa, S.; Shan, X. (2024). A multi-attention and depthwise separable convolution network for medical image segmentation. *Neurocomputing*, *564*, 126970. https://doi.org/10.1016/j.neucom.2023.126970.

[25] Chai, S.; et al. (2024). A Novel Adaptive Hypergraph Neural Network for Enhancing Medical Image Segmentation. *Lecture Notes in Computer Science*, 23–33. https://doi.org/10.1007/978-3-031-72114-4_3.

[26] Suara, S.; Jha, A.; Sinha, P.; Sekh, A.A. (2024). Is Grad-CAM Explainable in Medical Images?. *Lecture Notes in Computer Science*, 124–135. https://doi.org/10.1007/978-3-031-58181-6_11.

[27] Chen, L.; Cai, X.; Li, Z.; Xing, J.; Ai, J. (2024). Where is my attention? An explainable AI exploration in water detection from SAR imagery. *International Journal of Applied Earth Observation and Geoinformation*, *130*, 103878. https://doi.org/10.1016/j.jag.2024.103878.

[28] Kumar, D.; Porwal, S.; Malviya, R.; Sridhar, S.B. (2025). XAI Technique in Deep Learning–Based Medical Image Analysis. Explainable and Responsible Artificial Intelligence in Healthcare. Wiley: New Jersey, USA. pp. 191–215. https://doi.org/10.1002/9781394302444.ch8.

[29] Inigo, B.; Colliot, O.; Mitra, J. (2025). An intrinsically explainable approach to detecting vertebral compression fractures in CT scans via neurosymbolic modeling. Medical Imaging 2025: Image Processing, 101. SPIE. https://doi.org/10.1117/12.3047426.

[30] Gonzalez-Garcia, X.; Fumanal-Idocin, J.; Do Rio, J.M.N.; Bustince, H. (2024). A Rule-Based Approach for Interpretable Intensity-Modulated Radiation Therapy Treatment Selection. *IEEE International Conference on Fuzzy Systems*, 1–8. https://doi.org/10.1109/FUZZ-IEEE60900.2024.10612073.

[31] Singhal, A.; Agrawal, K.K.; Quezada, A.; Aguiñaga, A.R.; Jiménez, S.; Yadav, S.P. (2024). Explainable Artificial Intelligence (XAI) Model for Cancer Image Classification. *Computer Modeling in Engineering & Sciences*, *141*(1), 401–441. https://doi.org/10.32604/cmes.2024.051363.

[32] Reis, E.P.; et al. (2024). Automated abdominal CT contrast phase detection using an interpretable and open-source artificial intelligence algorithm. *European Radiology*, *34*(10), 6680–6687. https://doi.org/10.1007/s00330-024-10769-6.

[33] Wei, Y.; Tam, R.; Tang, X. (2023). MProtoNet: A Case-Based Interpretable Model for Brain Tumor Classification with 3D Multi-parametric Magnetic Resonance Imaging. *Proceedings of Machine Learning Research*, 1798–1812.

[34] Brás, C.; et al. (2025). Explainable AI for medical image analysis. In Trustworthy AI in Medical Imaging; Elsevier: Amsterdam, Netherlands, pp. 347–366. https://doi.org/10.1016/B978-0-44-323761-4.00028-6.

[35] Lamy, J.-B.; Sekar, B.; Guezennec, G.; Bouaud, J.; Séroussi, B. (2019). Explainable artificial intelligence for breast cancer: A visual case-based reasoning approach. *Artificial Intelligence in Medicine*, *94*, 42–53. https://doi.org/10.1016/j.artmed.2019.01.001.

[36] Bouzar-Benlabiod, L.; Harrar, K.; Yamoun, L.; Khodja, M.Y.; Akhloufi, M.A. (2023). A novel breast cancer detection architecture based on a CNN-CBR system for mammogram classification. *Computers in Biology and Medicine*, *163*, 107133. https://doi.org/10.1016/j.compbiomed.2023.107133.

[37] Saw, S.N.; Yan, Y.Y.; Ng, K.H. (2025). Current status and future directions of explainable artificial intelligence in medical imaging. *European Journal of Radiology*, *183*, 111884. https://doi.org/10.1016/j.ejrad.2024.111884.

[38] Kelly, L.; Masek, M.; Lam, C.P.; Abela, B.; Gupta, A. (2024). Evolving Visual Counterfactual Medical Imagery Explanations with Cooperative Co-evolution using Dynamic Decomposition. *Proceedings of the Genetic and Evolutionary Computation Conference Companion*, 291–294. ACM. https://doi.org/10.1145/3638530.3654224.

[39] Ilanchezian, I.; Kobak, D.; Faber, H.; Ziemssen, F.; Berens, P.; Ayhan, M.S. (2021). Interpretable Gender Classification from Retinal Fundus Images Using BagNets. *Lecture Notes in Computer Science*, 477–487. https://doi.org/10.1007/978-3-030-87199-4_45.

[40] Komorowski, P.; Baniecki, H.; Biecek, P. (2023). Towards Evaluating Explanations of Vision Transformers for Medical Imaging. *IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops*, 3726–3732. https://doi.org/10.1109/CVPRW59228.2023.00383.

[41] Guo, K.H.; Chaudhari, N.N.; Jafar, T.; Chowdhury, N.F.; Bogdan, P.; Irimia, A. (2024). Anatomic Interpretability in Neuroimage Deep Learning: Saliency Approaches for Typical Aging and Traumatic Brain Injury. *Neuroinformatics*, *22*(4), 591–606. https://doi.org/10.1007/s12021-024-09694-2.

[42] Bhati, D.; Neha, F.; Amiruzzaman, M. (2024). A Survey on Explainable Artificial Intelligence (XAI) Techniques for Visualizing Deep Learning Models in Medical Imaging. *Journal of Imaging*, *10*(10), 239. https://doi.org/10.3390/jimaging10100239.

[43] Ukwuoma, C.C.; et al. (2025). Enhancing histopathological medical image classification for Early cancer diagnosis using deep learning and explainable AI – LIME & SHAP. *Biomedical Signal Processing and Control*, *100*, 107014. https://doi.org/10.1016/j.bspc.2024.107014.

[44] Borys, K.; et al. (2023). Explainable AI in medical imaging: An overview for clinical practitioners – Saliency-based XAI approaches. *European Journal of Radiology*, *162*, 110787. https://doi.org/10.1016/j.ejrad.2023.110787.

[45] Corbat, L.; Nauval, M.; Henriet, J.; Lapayre, J.-C. (2020). A fusion method based on Deep Learning and Case-Based Reasoning which improves the resulting medical image segmentations. *Expert Systems with Applications*, *147*, 113200. https://doi.org/10.1016/j.eswa.2020.113200.

[46] Tiwari, R.G.; Misra, A.; Maheshwari, S.; Gautam, V.; Sharma, P.; Trivedi, N.K. (2025). Adaptive neuro-FUZZY inference system-fusion-deep belief network for brain tumor detection using MRI images with feature extraction. *Biomedical Signal Processing and Control*, *103*, 107387. https://doi.org/10.1016/j.bspc.2024.107387.

[47] Rabbah, J.; Ridouani, M.; Hassouni, L. (2025). Improving pneumonia diagnosis with high-accuracy CNN-Based chest X-ray image classification and integrated gradient. *Biomedical Signal Processing and Control*, *101*, 107239. https://doi.org/10.1016/j.bspc.2024.107239.

[48] Muscogiuri, G.; et al. (2022). Application of AI in cardiovascular multimodality imaging. *Heliyon*, *8*(10), e10872. https://doi.org/10.1016/j.heliyon.2022.e10872.

[49] Lanzafame, L.R.M.; et al. (2023). Artificial Intelligence in Cardiovascular CT and MR Imaging. *Life*, *13*(2), 507. https://doi.org/10.3390/life13020507.

[50] Cau, R.; et al. (2021). Potential Role of Artificial Intelligence in Cardiac Magnetic Resonance Imaging. *Journal of Thoracic Imaging*, *36*(3), 142–148. https://doi.org/10.1097/RTI.0000000000000584.

[51] Khanna, S.T.; Khatri, S.K.; Sharma, N.K. (2024). EXAIOC: Explainable AI for Oral Cancer Diagnosis and Prognosis - An Application-Centric Approach for Early Detection and Treatment

Planning. *Lecture Notes in Computer Science*, 223–237. https://doi.org/10.1007/978-3-031-70018-7_25.

[52] Martínez-Ramírez, J.M.; Carmona, C.; Ramírez-Expósito, M.J.; Martínez-Martos, J.M. (2025). Extracting Knowledge from Machine Learning Models to Diagnose Breast Cancer. *Life*, *15*(2), 211. https://doi.org/10.3390/life15020211.

[53] Hassan, M.R.; Hassan, M.M.; Rahman, M.A. (2024). Comparative Analysis of LIME and Shape Analysis Techniques in Prostate Cancer MRI Interpretation. International Conference on Engineering and Emerging Technologies, 1–6. IEEE. https://doi.org/10.1109/ICEET65156.2024.10913801.

[54] Yukta; Biswas, A.P.; Kashyap, S. (2024). Explainable AI for Healthcare Diagnosis in Renal Cancer. OPJU International Technology Conference (OTCON) on Smart Computing for Innovation and Advancement in Industry 4.0, 1–6. https://doi.org/10.1109/OTCON60325.2024.10687607.

[55] Aswin, R.; Kumar, H.A.; Varma, U.R. (2024). ResNet-Based Deep Learning Framework for Liver Cancer Detection with Explainable AI (XAI) Technique. 15th International Conference on Computing Communication and Networking Technologies, 1–6. https://doi.org/10.1109/ICCCNT61001.2024.10724461.

[56] Sinha, A.; Perti, A. (2024). Algorithm transparency and interpretability for AI-based medical imaging. IGI Global: Hershey, USA, 339–364. https://doi.org/10.4018/979-8-3693-5226-7.ch013.

[57] Chehri, A.; Ahmed, I.; Jeon, G. (2024). From Deep Learning to Interpretable and Explainable Deep Learning in Medical Image Computing: Balancing Innovation with Ethics and Responsibilities. *Procedia Computer Science*, *246*, 302–311. https://doi.org/10.1016/j.procs.2024.09.409.

[58] de C. Aranovich, T.; Matulionyte, R. (2023). Ensuring AI explainability in healthcare: problems and possible policy solutions. *Information & Communications Technology Law*, *32*(2), 259–275. https://doi.org/10.1080/13600834.2022.2146395.

[59] Mienye, I.D.; et al. (2024). A survey of explainable artificial intelligence in healthcare: Concepts, applications, and challenges. *Informatics in Medicine Unlocked*, *51*, 101587. https://doi.org/10.1016/j.imu.2024.101587.

[60] Grover, V.; Dogra, M. (2024). Challenges and Limitations of Explainable AI in Healthcare. *Book Chapter*, 72–85. https://doi.org/10.4018/979-8-3693-5468-1.ch005.

[61] McNamara, S.L.; Yi, P.H.; Lotter, W. (2024). The clinician-AI interface: intended use and explainability in FDA-cleared AI devices for medical image interpretation. *NPJ Digital Medicine*, *7*(1), 80. https://doi.org/10.1038/s41746-024-01080-1.

[62] Amann, J.; Bürger, V.K.; Livne, M.; Bui, C.K.T.; Madai, V.I. (2025). The fundamentals of AI ethics in medical imaging. In Trustworthy AI in Medical Imaging; Elsevier: Amsterdam, Netherlands, pp. 7–33. https://doi.org/10.1016/B978-0-44-323761-4.00010-9.

[63] Panigutti, C.; et al. (2023). Co-design of Human-centered, Explainable AI for Clinical Decision Support. *ACM Transactions on Interactive Intelligent Systems*, *13*(4), 1–35. https://doi.org/10.1145/3587271.

[64] Shafik, W.; Lakshmi, D. (2024). Explainable AI (EXAI) for Smart Healthcare Automation. In *Book Title*; IGI Global: Publisher Location, Country, pp. 289–316. https://doi.org/10.4018/979-8-3693-4439-2.ch012.

[65] Fontes, M.; De Almeida, J.D.S.; Cunha, A. (2024). Application of Example-Based Explainable Artificial Intelligence (XAI) for Analysis and Interpretation of Medical Imaging: A Systematic Review. *IEEE Access*, *12*, 26419–26427. https://doi.org/10.1109/ACCESS.2024.3367606.

[66] Shad, R.; Cunningham, J.P.; Ashley, E.A.; Langlotz, C.P.; Hiesinger, W. (2021). Designing clinically translatable artificial intelligence systems for high-dimensional medical imaging. *Nature Machine Intelligence*, *3*(11), 929–935. https://doi.org/10.1038/s42256-021-00399-8.

[67] Urvi; Sharma, P.; Goyal, K.; Sharma, S. (2025). Real-World Applications of Explainable AI in Healthcare. In *Explainable Artificial Intelligence in the Healthcare Industry*; Wiley: Hoboken, NJ, USA, pp. 451–466. https://doi.org/10.1002/9781394249312.ch20.

[68] Chaddad, A.; Hu, Y.; Wu, Y.; Wen, B.; Kateb, R. (2025). Generalizable and explainable deep learning for medical image computing: An overview. *Current Opinion in Biomedical Engineering*, *33*, 100567. https://doi.org/10.1016/j.cobme.2024.100567.

[69] Sinha, A.; Garcia, D.W.; Kumar, B.; Banerjee, P. (2023). Application of Big Data Analytics and Internet of Medical Things (IoMT) in Healthcare with View of Explainable Artificial Intelligence: A Survey. *Lecture Notes in Computer Science*, 129–163. https://doi.org/10.1007/978-3-031-08637-3_8.

[70] Linguraru, M.G.; et al. (2024). Clinical, Cultural, Computational, and Regulatory Considerations to Deploy AI in Radiology: Perspectives of RSNA and MICCAI Experts. *Radiology: Artificial Intelligence*, *6*(4). https://doi.org/10.1148/ryai.240225.

[71] Ennab, M.; Mcheick, H. (2025). A novel convolutional interpretability model for pixel-level interpretation of medical image classification through fusion of machine learning and fuzzy logic. *Smart Health*, *35*, 100535. https://doi.org/10.1016/j.smhl.2024.100535.