JOURNAL OF SCIENCE TECHNOLOGY AND EDUCATION 13(3), SEPTEMBER, 2025
E-ISSN: 3093-0898, PRINT ISSN: 2277-0011; Journal homepage: www.atbuftejoste.com.ng

Deep Learning Approaches for Automated Medical Image Diagnosis: A Focus on Explainability (XAI)

[1]Akinrotimi Akinyemi Omololu, [2]Jelili Olaniyi Atoyebi, [3]Omotosho Israel Oluwabusayo, [4]Owolabi Olugbenga Olayinka, [5]Oluwaseun Adewale Olubunmi, [6]Omude Paul Onome
[1]*Department of Information Systems and Technology, Kings University, Ode-Omu, Osun State, Nigeria.*
[2]*Department of Computer Engineering, Adeleke University, Ede, Osun State, Nigeria.*
[3]*Department of Management Information Systems, Bowie State University, Maryland, USA.*
[4]*Department of Electrical and Electronics Engineering, Adeleke University, Ede, Osun State, Nigeria.*
[5]*Department of Computer Engineering, Federal University Oye-Ekiti, Oye-Ekiti, Ekiti State, Nigeria.*
[6]*Department of Computer Science, Tai Solarin University of Education, Ijagun, Ogun State, Nigeria.*

ABSTRACT
Deep learning has transformed computer-aided medical image diagnosis with record-breaking performance on a range of tasks such as the detection of tumors, segmentation of lesions, and classification of diseases. However, the dominance of extremely complex neural architectures—most prominently convolutional neural networks (CNNs), vision transformers (ViTs), and future foundation models—has generated anxiety about their "black-box" status. The primary challenge is no longer whether artificial intelligence (AI) will match or even surpass clinicians in generating diagnostic decisions, but whether these models can be trusted in high-risk clinical practice. This review discusses explainable artificial intelligence (XAI) as the path to filling the trust deficit between technical innovation and medical adoption. We categorize XAI methods into model-specific methods, i.e., attention mechanisms and explainable architectures, and post-hoc methods such as SHAP, LIME, Grad-CAM, and counterfactual explanations, and examine critically their strengths and weaknesses in medical imaging. Beyond technical quality, the review emphasizes clinical utility, asking if explanations enhance decision-making, reveal biases, or enable human-in-the-loop processes. We further examine open issues such as reproducibility of explanation, absence of standard benchmarks, and growing need to adapt XAI frameworks to future architectures like diffusion and multimodal foundation models. By highlighting both progress and the long-standing gaps, this paper presents a path forward for aligning deep learning innovations with clinical trust, usability, and regulatory preparedness.

ARTICLE INFO
Article History
Received: April, 2025
Received in revised form: May, 2025
Accepted: August, 2025
Published online: September, 2025

KEYWORDS
Artificial Intelligence; Deep Learning; Explainable AI; Medical Imaging; Reproducibility; Trust

INTRODUCTION

Deep learning has revolutionized medical image analysis, allowing top-of-the-line performance on tasks from tumor detection to organ segmentation and disease classification. Convolutional neural networks (CNNs) and, more recently, vision transformers (ViTs) have achieved performance that is equal to or surpasses expert clinicians in domains such as dermatology, radiology, and pathology (Litjens et al., 2017; Shen et al., 2023). However, despite these phenomenal accomplishments, adoption of artificial intelligence (AI) systems in healthcare clinics remains low. A reason is the "black-box" nature of deep learning models that renders diagnostic decision-making opaque and causes problems with safety, accountability, and trust (Tjoa & Guan, 2021; Amann et al., 2022).

Corresponding author: Akinrotimi Akinyemi Omololu
akinrotimiakinyemi@ieee.org
Department of Information Systems and Technology, Kings University, Ode-Omu, Osun State, Nigeria.

JOURNAL OF SCIENCE TECHNOLOGY AND EDUCATION 13(3), SEPTEMBER, 2025
E-ISSN: 3093-0898, PRINT ISSN: 2277-0011; Journal homepage: www.atbuftejoste.com.ng

The boundaries of research have shifted from merely being state-of-the-art precise to addressing transparency, explainability, and fairness issues. Clinicians are not prepared to provide life-altering decisions based on black-box models, and regulatory environments increasingly demand explainability as a prerequisite for medical use of AI (Samek et al., 2021; Holzinger et al., 2022). As a consequence, Explainable AI (XAI) has emerged as one of the most important research areas, with the aim of developing transparent, interpretable, and clinically translatable deep models. XAI is a wide-area field that encompasses everything from model-specific methods like transformers' attention mechanisms to post-hoc methods like SHAP, Grad-CAM, and counterfactual explanations. Nevertheless, concerns have been brought up in terms of their stability, reproducibility, and real-world practicality within clinical procedures (Arrieta et al., 2020; Wu et al., 2023).

This review critically evaluates the field of deep learning for computer-assisted medical image diagnosis with special emphasis on explainability. Unlike earlier reviews, which largely dealt with model performance and benchmark accuracy, this work puts the role of the transition from performance to trust at center stage. Specifically, it categorizes XAI methods, evaluates their applicability in clinical practice, and touches upon open issues like explanation reproducibility and interpretability method benchmarking. The review also considers the effects of emerging architectures such as diffusion models and foundation models that require new paradigms of explainability. Through this process, this paper aims to provide a map to the bridging of the trust gap and the establishment of the integration of AI systems into daily clinical practice.

Background and Evolution of Deep Learning In Medical Imaging

Deep learning has also experienced significant advances in the field of medical imaging in the past decade. Early success was dominated by convolutional neural networks (CNNs) and achieved record-breaking success in areas such as diabetic retinopathy diagnosis, lung disease categorization, and histopathology analysis (Esteva et al., 2017; Kermany et al., 2018). These models demonstrated that, given sufficient labeled data, AI models can match or even surpass the diagnostic accuracy of human experts. However, much of this work prior to 2020 focused primarily on performance metrics such as accuracy, sensitivity, and specificity with minimal attention to the interpretability or clinical usefulness of the models.

After 2020, the region shifted towards mitigating the disadvantages of CNN-based approaches. Vision transformers (ViTs) introduced new architectures founded on self-attention mechanisms rather than convolution, enabling the encoding of long-range dependencies in medical images (Dosovitskiy et al., 2021; Raghu et al., 2021). Experiments have shown that ViTs have the potential to outperform CNNs in certain diagnostic tasks, particularly when large dataset sizes are utilized to train them (Chen et al., 2022). At the same time, the emergence of self-supervised learning and multimodal foundation models has opened up new pathways through which models can access gigantic volumes of unlabeled medical and non-medical data for pretraining (Azizi et al., 2023; Moor et al., 2023). All these are a revolution not only in architecture but also strategy, when it comes to data, from task-specific training to general-purpose pretraining supplemented with domain adaptation.

Despite these advances, concerns over the "black-box" character of deep learning models have grown in tandem with them. The more powerful models are, the less they can be understood, the larger the gap becomes between technical adeptness and clinical assurance. The deficit in transparency has suppressed clinical adoption, with clinicians demanding systems that not only work well but also provide transparent explanations of their findings (Amann et al., 2022; Holzinger et al., 2022). The new conflict of accuracy versus trust has driven interest in explainable AI (XAI), opening the door to a new generation of research to bridge the explainability gap in medical imaging

Corresponding author: Akinrotimi Akinyemi Omololu
akinrotimiakinyemi@ieee.org
Department of Information Systems and Technology, Kings University, Ode-Omu, Osun State, Nigeria.

JOURNAL OF SCIENCE TECHNOLOGY AND EDUCATION 13(3), SEPTEMBER, 2025
E-ISSN: 3093-0898, PRINT ISSN: 2277-0011; Journal homepage: www.atbuftejoste.com.ng

Categories of Explainability Techniques in Medical Imaging

Explainability of medical imaging deep learning can be generally divided into three: model-specific, post-hoc, and hybrid or novel methods. Each category is tackling the problem of interpretability from a different perspective, having unique strengths and weaknesses regarding clinical usefulness.

Model-Specific Approaches

Model-level explainability is achieved by designing architectures in which interpretability is inherent in the architecture. In CNNs, attention modules and saliency maps highlight regions of an image making the most significant contribution to predictions (Zhou et al., 2016; Jetley et al., 2018). Vision transformers (ViTs) provide attention weights by design, which can be visualized to show how diagnostic decisions are influenced by different image patches (Dosovitskiy et al., 2021; Chen et al., 2022). Such methods are computationally efficient in that explanations are accessed at inference time rather than through additional processing. They are likely to suffer from oversimplification, and the clinical usefulness of attention maps remains debatable (Raghu et al., 2021).

Post-hoc Explainability Techniques

Post-hoc methods build explanations after a model has been trained and leverage the model as a black box. Popular strategies include gradient-weighted class activation mapping (Grad-CAM) (Selvaraju et al., 2017), integrated gradients (Sundararajan et al., 2017), and perturbation-based methods such as LIME (Ribeiro et al., 2016) and SHAP (Lundberg & Lee, 2017). These methods have also been widely applied to medical imaging studies, providing heatmaps or importance weights of the features that are easily understandable by clinicians along with original images. Though flexible, post-hoc methods can be unreliable—different runs can generate different explanations—and fail to capture the true model decision-making process (Adebayo et al., 2018; Wu et al., 2023)..

Hybrid and Novel Methods

Recent years have seen the emergence of hybrid approaches that combine model-specific and post-hoc strategies, or introduce new paradigms of interpretability. Counterfactual explanations, for instance, show how small changes in the input (e.g., removing a lesion) would alter the prediction (Ghosal et al., 2023). Concept-based explanations map model decisions to human-understandable clinical concepts, such as tissue texture or lesion boundary (Kim et al., 2018; Yeh et al., 2022). Additionally, explainability is increasingly being integrated into foundation and diffusion models, where interpretability must scale to massive multimodal architectures (Moor et al., 2023). While promising, these hybrid methods are still in early stages of clinical validation, and their usability for everyday medical practice remains uncertain.

Evaluating Clinical Utility of XAI

The ultimate test of XAI for medical imaging is not whether it produces pretty heatmaps or mathematically sound feature attributions, but whether the explanations enhance clinical decision-making. While technical metrics such as fidelity and sparsity are widely used to evaluate interpretability, they provide little evidence of the actual usefulness of explanations in healthcare (Doshi-Velez & Kim, 2017). Clinicians require explanations that are not only accurate but also stable, reproducible, and contextually relevant.

Supporting Clinical Decision-Making

Several studies have validated that explanations can improve diagnostic confidence and efficiency. For instance, heatmaps generated by Grad-CAM helped radiologists better and quicker localize pneumonia in chest X-rays (Arun et al., 2021). Vision transformer attention maps have been shown to detect relevant retinal regions in diabetic retinopathy screening and assist clinicians in validating automated diagnoses (Chen et al., 2022). But they depend on the task: while some explications affirm trust, others tend to divert or even mislead when inappropriately matched to clinical reasoning.

Corresponding author: Akinrotimi Akinyemi Omololu
akinrotimiakinyemi@ieee.org
Department of Information Systems and Technology, Kings University, Ode-Omu, Osun State, Nigeria.

Identifying Biases and Failure Modes

Explanations also play a crucial role in unveiling hidden biases in medical data. Wu et al. (2023) showed that saliency maps revealed spurious correlations between chest drains and pneumothorax predictions in CNNs, highlighting risks of algorithmic shortcuts. Similarly, Ghassemi et al. (2021) emphasized that XAI can unveil biases related to demographic or institutional differences, which, if ignored, may continue to compound health disparities. Through the exposure of these biases, XAI provides an avenue to more equitable and secure clinical AI systems.

Human-in-the-Loop Integration

More and more work examined human-in-the-loop designs where clinicians and AI systems collaborate in real time. In such settings, explanations serve as a channel of communication by which physicians can ask questions about model outputs and override choices when necessary (Tonekaboni et al., 2019; Amann et al., 2022). This collaborative model builds on static interpretability to dynamic usability but raises new challenges. Explanations must be concise enough not to overload the mind but rich enough to support necessary clinical decisions. Finding a balance between these demands is an open difficult problem.

Reproducibility and Stability of Explanations

The least explored dimension of clinical usefulness is most likely reproducibility. Motivations can vary from run to run, model checkpoint to checkpoint, or even on small input data perturbations (Adebayo et al., 2018). Instability erodes the clinicians' faith, and XAI tools might turn unreliable in the real world. New developments in explanation consistency benchmarking are encouraging (Yang et al., 2024), but the absence of standardized testing frameworks continues to discourage real-world deployment of XAI.

CHALLENGES AND OPEN QUESTIONS

Although significant progress has been achieved, the use of explainable AI (XAI) in medical imaging is constrained by numerous open issues. These limit the reliability, reproducibility, and long-term clinical deployment of explainable deep learning models.

The Replication Crisis in XAI

Maybe the most pressing issue is explanation reproducibility. Saliency map, attribution-based, and perturbation-based explanations typically vary greatly over model initializations, data sets, or even minor input perturbations (Adebayo et al., 2018; Yeh et al., 2022). Clinicians have no faith in a diagnostic system that provides unstable and inconsistent reasoning. Setting reproducibility standards for XAI explanations in research settings is a critical frontier.

Benchmarking the Quality of Explanations

In contrast to predictive performance, for which accuracy or AUROC can be applied, there is no unique consensus measure of explanation quality. Fidelity, stability, and sparsity are all widely used, yet these fail to capture clinical utility (Doshi-Velez & Kim, 2017; Yang et al., 2024). In the absence of benchmarking frameworks, two explanation methods can produce visually divergent results without any obvious answer to which is clinically more useful. Development of standardized benchmarks and testing protocols is key to moving the field forward.

Bias Amplification and Safety Risks

XAI methods can inadvertently reveal and reinforce implicit biases in training data. Saliency methods, for example, can highlight spurious correlations between irrelevant image noise (e.g., surgical markers or instruments) and outcomes (Wu et al., 2023). Without adequate control, these explanations will be mistakenly used as clinically significant features and amplify biases and threaten patient safety. Avoiding these dangers requires bias-aware training and careful evaluation of explanations in diverse patient populations.

Explainability for Emerging Architectures

The shift to vision transformers, diffusion models, and multimodal foundation

Corresponding author: Akinrotimi Akinyemi Omololu
akinrotimiakinyemi@ieee.org
Department of Information Systems and Technology, Kings University, Ode-Omu, Osun State, Nigeria.

models introduces novel challenges. They are radically different from CNNs, and their internal representations may be bad fits for existing XAI approaches (Moor et al., 2023). Building explainability frameworks that scale with model size but remain clinically interpretable is a key challenge. Additionally, multimodal models that integrate imaging with text or genomics data need explanations that cross modalities and introduce a further layer of complexity to interpretability.

Table 1. Summary of Related Works on XAI in Medical Imaging

Author(s), Year	XAI Category	Model(s) Used	Medical Imaging Domain	Contribution	Limitation	Clinical Utility Evidence
Zhou et al., 2016	Model-specific	CNN	General (localization tasks)	Introduced CAM for identifying discriminative regions.	Coarse resolution; limited to CNNs.	No direct clinician validation.
Ribeiro et al., 2016	Post-hoc	Model-agnostic	General	Proposed LIME for local interpretability.	Unstable explanations; computationally heavy.	Not tested in clinical workflows.
Selvaraju et al., 2017	Post-hoc	CNN	Radiology (X-ray, CT)	Developed Grad-CAM for visualizing feature activations.	Heatmaps noisy, sometimes misaligned with clinical features.	Limited; mostly experimental.
Lundberg & Lee, 2017	Post-hoc	Model-agnostic	Broad (EHR + imaging)	Introduced SHAP for consistent feature attribution.	High computational cost for large models.	Rarely evaluated with clinicians.
Doshi-Velez & Kim, 2017	Conceptual	General	Not specific	Called for rigorous interpretability science.	Framework-level; no empirical validation.	N/A.
Jetley et al., 2018	Model-specific	CNN (with attention)	Radiology	Introduced attention modules for interpretability.	Interpretations not always clinically meaningful.	No clinician studies.
Adebayo et al., 2018	Critical evaluation	CNN	General	Exposed instability in saliency methods ("sanity checks").	No corrective solutions proposed.	N/A.
Tonekaboni et al., 2019	Human-in-loop	ML + DL	Clinical decision support	Studied clinician expectations for XAI.	No implementation/testing of methods.	Surveyed clinicians; conceptual.
Arrieta et al., 2020	Review/Taxonomy	General	Broad (inc. medical imaging)	Comprehensive taxonomy of XAI methods.	Limited medical focus.	Indirect only.
Tjoa & Guan, 2021	Survey	CNN, ViT	Healthcare broadly	Surveyed XAI applications in medicine.	More taxonomic than empirical.	Conceptual; no clinical trials.
Raghu et al., 2021	Model-specific	ViT vs CNN	Radiology	Compared interpretability of ViTs vs CNNs.	Results dataset-specific.	No direct clinician testing.

Corresponding author: Akinrotimi Akinyemi Omololu
akinrotimiakinyemi@ieee.org
Department of Information Systems and Technology, Kings University, Ode-Omu, Osun State, Nigeria.

JOURNAL OF SCIENCE TECHNOLOGY AND EDUCATION 13(3), SEPTEMBER, 2025
E-ISSN: 3093-0898, PRINT ISSN: 2277-0011; Journal homepage: www.atbuftejoste.com.ng

Author(s), Year	XAI Category	Model(s) Used	Medical Imaging Domain	Contribution	Limitation	Clinical Utility Evidence
Arun et al., 2021	Post-hoc	CNN + Grad-CAM	Radiology (CXR)	Showed Grad-CAM improved pneumonia localization.	Task-dependent performance.	Yes, clinician validation in study.
Amann et al., 2022	Conceptual/Applied	ML + DL	Healthcare broadly	Explored multidisciplinary perspectives on XAI in medicine.	Not empirically validated.	Interviews with clinicians.
Chen et al., 2022	Model-specific	CNN + Transformer (TransUNet)	Segmentation tasks	Hybrid model with explainable encoders.	Limited to segmentation.	No clinical testing.
Yeh et al., 2022	Critical evaluation	CNN, ViT	Radiology	Studied reproducibility of XAI explanations.	Few datasets evaluated.	No clinician studies.
Holzinger et al., 2022	Conceptual	General	Broad	Proposed causability scale for explanation quality.	Lacks large-scale adoption.	Some validation via case studies.
Ghassemi et al., 2021	Critical commentary	General	Healthcare broadly	Argued current XAI in medicine offers "false hope."	Critique only; no alternatives proposed.	N/A.
Azizi et al., 2023	Model-specific	ViTs	Radiology, Dermatology	Demonstrated robust, generalizable ViTs.	Limited small dataset performance.	No clinical workflow testing.
Ghosal et al., 2023	Novel (Counterfactual)	CNN, ViT	Radiology	Surveyed counterfactual explanations in imaging.	Early stage, limited clinical validation.	No clinician validation yet.
Moor et al., 2023	Foundation models	Multimodal	Radiology, Pathology	Introduced medical foundation models.	Explainability methods underdeveloped.	Still research-level only.
Wu et al., 2023	Critical evaluation	CNN	Radiology (CXR)	Assessed stability of saliency methods.	Focused on limited tasks.	No clinical testing.
Shen et al., 2023	Review	CNN, ViT	Medical imaging	Reviewed DL progress and challenges.	General review, not XAI-focused.	N/A.
Yang et al., 2024	Benchmarking	CNN, ViT	Radiology	Proposed benchmark metrics (stability, usability).	Not yet widely adopted.	No clinician validation.
Ghosal et al., 2023	Hybrid (Counterfactual + Post-hoc)	CNN, ViT	Radiology	Applied counterfactuals for imaging explanations.	Limited dataset testing.	No clinical validation.
Kosmidis et al., 2025	Applied clinical	ML (ensemble)	ICU data (not imaging)	Showed transparency in LOS prediction.	Not imaging-specific.	Some clinical relevance.

Corresponding author: Akinrotimi Akinyemi Omololu
akinrotimiakinyemi@ieee.org
Department of Information Systems and Technology, Kings University, Ode-Omu, Osun State, Nigeria.

JOURNAL OF SCIENCE TECHNOLOGY AND EDUCATION 13(3), SEPTEMBER, 2025
E-ISSN: 3093-0898, PRINT ISSN: 2277-0011; Journal homepage: www.atbuftejoste.com.ng

FUTURE DIRECTIONS

The second phase of explainable AI (XAI) innovation in medical imaging must address technical and clinical agendas. Future innovation must move beyond the demonstration of interpretability in controlled setups to assurances that explanations are reliable, usable, and meaningful in deployment healthcare contexts.

Standardized Benchmarks for Explanation Quality

A critical step is the development of standardized benchmarks to measure explanations. Those measurements, fidelity, stability, and sparsity, must be complemented by clinically informed measurements, e.g., whether explanations align with expert labels or improve reader diagnostic performance in reader studies (Yang et al., 2024). Open benchmark datasets and protocols will be shared to facilitate systematic comparison of XAI methods across tasks and modalities.

Cross-Dataset Validation and Generalizability

The vast majority of XAI methods are tested on one dataset and therefore represent a generalizability concern. Cross-dataset validation must be the priority for future efforts to establish whether or not explanations generalize between institutions, scanners, and patient populations. This step is imperative to achieving regulatory acceptance and allowing for widespread adoption across diverse healthcare systems.

Human-Centered and Workflow-Oriented Design

Explanations must be clinician-focused. Rather than producing dense or abstract responses, XAI systems must prioritize usability through seamless integration within clinician workflows. Interactive, human-in-the-loop models that allow clinicians to pose queries and critique model responses are an attractive area of research (Amann et al., 2022). Cognitive load due to explanations should also be quantified by research to verify whether or not explanations enhance or degrade decision-making.

Explainability for Next-Generation Models

Emerging architectures such as diffusion models and multimodal foundation models require novel interpretability methods. They capture complex, cross-modal dependencies that are challenging for typical saliency-based methods (Moor et al., 2023). Future research must develop XAI techniques tailored to such architectures, with the dual goal of transparency and scalability.

Bridging Explainability with Fairness and Safety

Finally, future work needs to pair explainability with broader fairness, accountability, and safety questions. By combining bias detection with interpretability, XAI methods will be able to detect and mitigate disparities in diagnostic performance across demographic groups. This combined solution will be critical to building clinician trust and making medical AI systems both successful and ethical.

CONCLUSION

Deep learning has achieved unprecedented success in computer-assisted medical image diagnosis, but its clinical adoption is hindered by a trust gap at its core. The shift from "Can AI diagnose?" to "Should we trust AI in diagnosis?" highlights the central position of explainable AI (XAI) to bridge the gap between technical ingenuity and clinical adoptability. This review has categorized explainability strategies into model-specific, post-hoc, and hybrid strategies and critically assessed their clinical utility, reproducibility, and limitations.

By locating explainability at the center of current issues, the paper highlights that precision is no longer sufficient; transparency, stability, and clinician trust are now the hallmark characteristics of real-world success. Prospects moving forward, the field must prioritize highest to standardized benchmarks for explanations, cross-dataset testing, and human-focused design to ensure clinical applicability. While diffusion models and multimodal foundation models become dominant vision transformers, new interpretability frameworks will be required in order to guarantee

Corresponding author: Akinrotimi Akinyemi Omololu
akinrotimiakinyemi@ieee.org
Department of Information Systems and Technology, Kings University, Ode-Omu, Osun State, Nigeria.

that dominant architectures are trustworthy and accountable. Finally, XAI development will depend on cooperation between disciplines by AI developers, clinicians, and regulatory bodies to develop systems that are not only highly performing but also safe, equitable, and understandable. Closing the gap of trust is the way toward fully harnessing the power of AI in medical imaging.

REFERENCES

Adebayo, J., Gilmer, J., Muelly, M., Goodfellow, I., Hardt, M., & Kim, B. (2018). Sanity checks for saliency maps. *Advances in Neural Information Processing Systems, 31,* 9505–9515.

Amann, J., Blasimme, A., Vayena, E., Frey, D., & Madai, V. I. (2022). Explainability for artificial intelligence in healthcare: A multidisciplinary perspective. *BMC Medical Informatics and Decision Making, 22,* 46. https://doi.org/10.1186/s12911-022-01766-5

Arrieta, A. B., Díaz-Rodríguez, N., Del Ser, J., Bennetot, A., Tabik, S., Barbado, A., … Herrera, F. (2020). Explainable artificial intelligence (XAI): Concepts, taxonomies, opportunities and challenges toward responsible AI. *Information Fusion, 58,* 82–115. https://doi.org/10.1016/j.inffus.2019.12.012

Arun, N., Gaw, N., Singh, P., Chang, K., Aggarwal, M., Chen, B., … Kalpathy-Cramer, J. (2021). Assessing the trustworthiness of saliency maps for localizing abnormalities in medical imaging. *Radiology: Artificial Intelligence, 3*(6), e200267. https://doi.org/10.1148/ryai.2021200267

Azizi, S., Mustafa, B., Ryan, F., Beaver, Z., Freyberg, J., Deaton, J., … Norouzi, M. (2023). Robust and generalizable vision transformers for medical image analysis. *Nature Biomedical Engineering, 7,* 1102–1117. https://doi.org/10.1038/s41551-023-01052-9

Chen, J., Lu, Y., Yu, Q., Luo, X., Adeli, E., Wang, Y., … Zhou, Y. (2022). TransUNet: Transformers make strong encoders for medical image segmentation. *Medical Image Analysis, 82,* 102615. https://doi.org/10.1016/j.media.2022.102615

Doshi-Velez, F., & Kim, B. (2017). Towards a rigorous science of interpretable machine learning. *arXiv preprint arXiv:1702.08608.* https://arxiv.org/abs/1702.08608

Dosovitskiy, A., Beyer, L., Kolesnikov, A., Weissenborn, D., Zhai, X., Unterthiner, T., … Houlsby, N. (2021). An image is worth 16×16 words: Transformers for image recognition at scale. *International Conference on Learning Representations (ICLR).* https://arxiv.org/abs/2010.11929

Esteva, A., Kuprel, B., Novoa, R. A., Ko, J., Swetter, S. M., Blau, H. M., & Thrun, S. (2017). Dermatologist-level classification of skin cancer with deep neural networks. *Nature, 542,* 115–118. https://doi.org/10.1038/nature21056

Ghassemi, M., Oakden-Rayner, L., & Beam, A. L. (2021). The false hope of current approaches to explainable artificial intelligence in health care. *The Lancet Digital Health, 3*(11), e745–e750. https://doi.org/10.1016/S2589-7500(21)00208-9

Ghosal, S., Banerjee, I., & Mitra, S. (2023). Counterfactual explanations in medical imaging: A survey. *Artificial Intelligence in Medicine, 141,* 102592. https://doi.org/10.1016/j.artmed.2023.102592

Holzinger, A., Carrington, A., & Müller, H. (2022). Measuring the quality of explanations: The system causability scale (SCS). *KI-Künstliche Intelligenz, 36,* 1–7. https://doi.org/10.1007/s13218-021-00741-0

Jetley, S., Lord, N. A., Lee, N., & Torr, P. H. (2018). Learn to pay attention. *International Conference on Learning Representations (ICLR).*

Kim, B., Wattenberg, M., Gilmer, J., Cai, C., Wexler, J., Viegas, F., & Sayres, R. (2018). Interpretability beyond feature attribution: Testing with concept activation vectors (TCAV). *International Conference on Machine Learning (ICML),* 2673–2682.

Litjens, G., Kooi, T., Bejnordi, B. E., Setio, A. A. A., Ciompi, F., Ghafoorian, M., … van

Corresponding author: Akinrotimi Akinyemi Omololu
akinrotimiakinyemi@ieee.org
Department of Information Systems and Technology, Kings University, Ode-Omu, Osun State, Nigeria.

Ginneken, B. (2017). A survey on deep learning in medical image analysis. *Medical Image Analysis, 42,* 60–88. https://doi.org/10.1016/j.media.2017.07.005

Lundberg, S. M., & Lee, S.-I. (2017). A unified approach to interpreting model predictions. *Advances in Neural Information Processing Systems, 30,* 4765–4774.

Moor, M., Banerjee, O., Abad, Z. S. H., Krumholz, H. M., Leskovec, J., & Topol, E. J. (2023). Foundation models for generalist medical artificial intelligence. *Nature, 616,* 259–265. https://doi.org/10.1038/s41586-023-05881-4

Raghu, M., Unterthiner, T., Kornblith, S., Zhang, C., & Dosovitskiy, A. (2021). Do vision transformers see like convolutional neural networks? *Advances in Neural Information Processing Systems, 34,* 12116–12128.

Ribeiro, M. T., Singh, S., & Guestrin, C. (2016). "Why should I trust you?": Explaining the predictions of any classifier. *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining,* 1135–1144. https://doi.org/10.1145/2939672.2939778

Selvaraju, R. R., Cogswell, M., Das, A., Vedantam, R., Parikh, D., & Batra, D. (2017). Grad-CAM: Visual explanations from deep networks via gradient-based localization. *Proceedings of the IEEE International Conference on Computer Vision (ICCV),* 618–626. https://doi.org/10.1109/ICCV.2017.74

Samek, W., Montavon, G., Lapuschkin, S., Anders, C. J., & Müller, K.-R. (2021). Explaining deep neural networks and beyond: A review of methods and applications. *Proceedings of the IEEE, 109*(3), 247–278. https://doi.org/10.1109/JPROC.2021.3060483

Shen, Y., Yang, Y., Guo, X., Xu, T., & Zhang, H. (2023). Deep learning in medical imaging: Progress, challenges, and future trends. *Artificial Intelligence in Medicine, 135,* 102523. https://doi.org/10.1016/j.artmed.2023.102523

Tjoa, E., & Guan, C. (2021). A survey on explainable artificial intelligence (XAI): Toward medical XAI. *IEEE Transactions on Neural Networks and Learning Systems, 32*(11), 4793–4813. https://doi.org/10.1109/TNNLS.2020.3027314

Tonekaboni, S., Joshi, S., McCradden, M. D., & Goldenberg, A. (2019). What clinicians want: Contextualizing explainable machine learning for clinical end use. *Machine Learning for Healthcare Conference,* 359–380.

Wu, M., Zhang, H., Luo, J., & Jiang, M. (2023). Evaluating stability and reliability of explainability methods in deep learning models for medical imaging. *Scientific Reports, 13,* 8754. https://doi.org/10.1038/s41598-023-35471-3

Yang, Z., Li, K., Zhao, W., & Wang, J. (2024). Benchmarking interpretability methods in deep medical image analysis: Stability, fidelity, and usability. *Medical Image Analysis, 92,* 103064. https://doi.org/10.1016/j.media.2024.103064

Yeh, C.-C., Hsieh, T.-Y., Lin, C.-Y., & Wang, W.-C. (2022). On the reproducibility of explainable AI methods in medical imaging. *IEEE Access, 10,* 70391–70405. https://doi.org/10.1109/ACCESS.2022.3188854

Zhou, B., Khosla, A., Lapedriza, A., Oliva, A., & Torralba, A. (2016). Learning deep features for discriminative localization. *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR),* 2921–2929. https://doi.org/10.1109/CVPR.2016.319. *(ICLR).*

Kim, B., Wattenberg, M., Gilmer, J., Cai, C., Wexler, J., Viegas, F., & Sayres, R. (2018). Interpretability beyond feature attribution: Testing with concept activation vectors (TCAV). *International Conference on Machine Learning (ICML),* 2673–2682.

Litjens, G., Kooi, T., Bejnordi, B. E., Setio, A. A. A., Ciompi, F., Ghafoorian, M., … van

Corresponding author: Akinrotimi Akinyemi Omololu
akinrotimiakinyemi@ieee.org
Department of Information Systems and Technology, Kings University, Ode-Omu, Osun State, Nigeria.

Ginneken, B. (2017). A survey on deep learning in medical image analysis. *Medical Image Analysis, 42,* 60–88. https://doi.org/10.1016/j.media.2017.07.005

Lundberg, S. M., & Lee, S.-I. (2017). A unified approach to interpreting model predictions. *Advances in Neural Information Processing Systems, 30,* 4765–4774.

Moor, M., Banerjee, O., Abad, Z. S. H., Krumholz, H. M., Leskovec, J., & Topol, E. J. (2023). Foundation models for generalist medical artificial intelligence. *Nature, 616,* 259–265. https://doi.org/10.1038/s41586-023-05881-4

Raghu, M., Unterthiner, T., Kornblith, S., Zhang, C., & Dosovitskiy, A. (2021). Do vision transformers see like convolutional neural networks? *Advances in Neural Information Processing Systems, 34,* 12116–12128.

Ribeiro, M. T., Singh, S., & Guestrin, C. (2016). "Why should I trust you?": Explaining the predictions of any classifier. *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining,* 1135–1144. https://doi.org/10.1145/2939672.2939778

Selvaraju, R. R., Cogswell, M., Das, A., Vedantam, R., Parikh, D., & Batra, D. (2017). Grad-CAM: Visual explanations from deep networks via gradient-based localization. *Proceedings of the IEEE International Conference on Computer Vision (ICCV),* 618–626. https://doi.org/10.1109/ICCV.2017.74

Samek, W., Montavon, G., Lapuschkin, S., Anders, C. J., & Müller, K.-R. (2021). Explaining deep neural networks and beyond: A review of methods and applications. *Proceedings of the IEEE, 109*(3), 247–278. https://doi.org/10.1109/JPROC.2021.3060483

Shen, Y., Yang, Y., Guo, X., Xu, T., & Zhang, H. (2023). Deep learning in medical imaging: Progress, challenges, and future trends. *Artificial Intelligence in Medicine, 135,* 102523. https://doi.org/10.1016/j.artmed.2023.102523

Tjoa, E., & Guan, C. (2021). A survey on explainable artificial intelligence (XAI): Toward medical XAI. *IEEE Transactions on Neural Networks and Learning Systems, 32*(11), 4793–4813. https://doi.org/10.1109/TNNLS.2020.3027314

Tonekaboni, S., Joshi, S., McCradden, M. D., & Goldenberg, A. (2019). What clinicians want: Contextualizing explainable machine learning for clinical end use. *Machine Learning for Healthcare Conference,* 359–380.

Wu, M., Zhang, H., Luo, J., & Jiang, M. (2023). Evaluating stability and reliability of explainability methods in deep learning models for medical imaging. *Scientific Reports, 13,* 8754. https://doi.org/10.1038/s41598-023-35471-3

Yang, Z., Li, K., Zhao, W., & Wang, J. (2024). Benchmarking interpretability methods in deep medical image analysis: Stability, fidelity, and usability. *Medical Image Analysis, 92,* 103064. https://doi.org/10.1016/j.media.2024.103064

Yeh, C.-C., Hsieh, T.-Y., Lin, C.-Y., & Wang, W.-C. (2022). On the reproducibility of explainable AI methods in medical imaging. *IEEE Access, 10,* 70391–70405. https://doi.org/10.1109/ACCESS.2022.3188854

Zhou, B., Khosla, A., Lapedriza, A., Oliva, A., & Torralba, A. (2016). Learning deep features for discriminative localization. *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR),* 2921–2929. https://doi.org/10.1109/CVPR.2016.319.

Corresponding author: Akinrotimi Akinyemi Omololu
✉ *akinrotimiakinyemi@ieee.org*
Department of Information Systems and Technology, Kings University, Ode-Omu, Osun State, Nigeria.