Explainable AI: Current Status and Future Potential in Radiology

Dr. Ashok Singh

Asst. Prof., Computer Science Dept. Shiksha Snatak Mahavidyalaya, Raipur (C.G.) India.

ABSTRACT

Explainable Artificial Intelligence (XAI) is rapidly evolving and holds significant promise for revolutionizing radiology practice. This paper presents an overview of the current status and future potential of XAI in radiology. We discuss the importance of XAI in enhancing transparency, interpretability, and trustworthiness in AI-driven decision-making processes within the radiological domain. Various XAI techniques, including visual explanations, textual descriptions, and example-based explanations, are examined in the context of their applications and limitations in radiology. Moreover, we explore emerging trends such as the integration of biological explanations and causal relationships into XAI frameworks. The paper underscores the need for collaborative efforts to define evaluation criteria and prioritize aspects crucial for the development of personalized XAI solutions tailored to the needs of clinicians, radiologists, and patients, while adhering to regulatory standards. By actively engaging in the direction of XAI, the radiology community can shape a future where AI-driven technologies optimize diagnostic accuracy, improve patient outcomes, and facilitate informed clinical decision-making.

Keywords: - XAI, Radiology, DARPA, AI, Techniques.

I INTRODUCTION

There has been a notable rise in the integration of artificial intelligence (AI) into critical decision-making processes that significantly impact human lives across various domains, such as the criminal justice system, autonomous vehicles, food safety, and radiology. In the realm of radiology, the prevailing method for AI implementation is through deep learning—a technique that employs neural networks comprising numerous interconnected layers to decipher complex patterns and relationships within data. However, the intricacies of these neural networks, characterized by their nonlinear dynamics and multilayered structures, often render them opaque to human understanding. Consequently, deep learning models are often referred to as "black boxes" due to the challenge of comprehending the internal mechanisms driving their decisions.

This opacity raises concerns regarding the potential presence of unnoticed biases within these black box systems, which could yield significant repercussions in scenarios involving high-stakes decision-making. The inability to fully elucidate how these neural networks arrive at their conclusions underscores the importance of addressing transparency and accountability issues inherent in AI systems, particularly when they wield considerable influence over human outcomes. (1-2)

The burgeoning need to enhance our comprehension of the opaque nature of deep learning has catalyzed the development of methods aimed at rendering AI systems more transparent and interpretable. These methods, commonly termed explainable artificial intelligence (XAI), have garnered increasing attention and demand within the AI research and development community. By employing various techniques and approaches, XAI endeavors to demystify the inner workings of deep learning models, allowing for a more comprehensive understanding of how they arrive at their decisions and predictions.

Recognizing the critical importance of transparency and interpretability in AI systems, XAI methodologies aim to bridge the gap between the complex computations of deep neural networks and human comprehension. By providing insights into the factors and features influencing model outputs, XAI not only fosters trust and accountability but also facilitates the identification and mitigation of potential biases and errors inherent in AI-driven decision-making processes.

As the demand for trustworthy and ethically sound AI continues to escalate, the development and integration of explainable artificial intelligence represent pivotal strides toward fostering transparency, accountability, and responsible deployment of AI technologies across diverse domains and applications. (3).

Several prominent initiatives in the realm of explainable artificial intelligence (XAI) have emerged, among which notable examples include efforts by the United States Defense Advanced Research Projects Agency (DARPA) and the Association for Computing Machinery's (ACM) conferences on Fairness, Accountability, and Transparency (ACM FAccT).

DARPA has been at the forefront of advancing XAI research through its programs aimed at enhancing the transparency and interpretability of AI systems, particularly within the context of defense and national security applications. These initiatives encompass a diverse array of projects and collaborations aimed at developing innovative approaches and tools for elucidating the decision-making processes of complex AI models.

Similarly, the ACM's conferences on Fairness, Accountability, and Transparency (ACM FAccT) serve as crucial platforms for interdisciplinary dialogue and research dissemination on the ethical, societal, and technical dimensions of AI transparency and accountability. These conferences bring together researchers, practitioners, policymakers, and stakeholders from across the globe to explore cutting-edge developments, share insights, and foster collaborations aimed at advancing the state-of-the-art in XAI.

By fostering collaboration, knowledge-sharing, and interdisciplinary engagement, initiatives spearheaded by DARPA and ACM FAccT play instrumental roles in driving forward the field of explainable artificial intelligence, thereby promoting transparency, accountability, and responsible innovation in AI systems and applications. (4-5).

Within the domain of medical imaging, an annual workshop titled Interpretability of Machine Intelligence in Medical Image Computing (iMIMIC) convenes as part of the International Conference on Medical Image Computing and Computer Assisted Intervention (MICCAI). This workshop serves as a focal point for researchers, practitioners, and experts in the field to explore and address challenges related to the interpretability and transparency of machine learning algorithms in medical image analysis.

By providing a dedicated forum for discussion, collaboration, and knowledge exchange, iMIMIC plays a pivotal role in advancing the understanding and development of interpretable machine learning techniques tailored to the specific demands and nuances of medical imaging applications. Through presentations, panel discussions, and interactive sessions, participants engage in cutting-edge research, share insights, and explore innovative methodologies aimed at enhancing the interpretability, reliability, and clinical relevance of AI-driven medical imaging systems.

The iMIMIC workshop at MICCAI underscores the growing recognition of the importance of transparency, interpretability, and accountability in the deployment of machine learning algorithms within clinical settings. By fostering interdisciplinary dialogue and collaboration among researchers, clinicians, and industry stakeholders, iMIMIC contributes to the continued evolution and refinement of machine learning techniques for medical image analysis, ultimately striving to enhance patient care, diagnostic accuracy, and clinical decision-making processes. (6)

II CURRENT XAI STATUS

In current XAI techniques applied to radiology, explanations typically manifest in visual, textual, example-based formats, or a combination thereof. Visual explanations often present a "heat map" or "saliency map," elucidating the areas where the algorithm primarily based its decision. Visual explanations stand out as the most prevalent XAI technique employed in radiology presently.

Textual explanations furnish descriptive narratives, ranging from straightforward annotations like "hyper intense lesion" to comprehensive medical reports encapsulating various aspects of the diagnosis.

Example-based explanations offer relevant instances to illustrate how a neural network arrived at its decision, akin to how a radiologist draws insights from past cases to assess the current one.

These diverse XAI techniques in radiology cater to different preferences and requirements, each offering unique insights into the decision-making process of AI systems applied in medical imaging. (7)

Post hoc XAI methods, which provide explanations after a neural network has been trained, offer several advantages. These techniques are often open source and user-friendly, particularly within frameworks like captum.ai. Additionally, post hoc XAI is typically model agnostic, enabling it to generate explanations regardless of the underlying algorithm, thereby facilitating the provision of explanations for neural networks already deployed in clinical settings.

Despite these advantages, post hoc XAI methods also entail notable disadvantages. They may exhibit unexpected behavior, and not all techniques demonstrate high validity, defined as the accuracy and relevance of the explanation to the end user's expectations. Concerns regarding robustness further underscore the need for caution when relying solely on post hoc XAI methods.

A practical approach to mitigate these disadvantages involves leveraging multiple post hoc XAI techniques and evaluating the consistency and reliability of their explanations. By cross-referencing explanations from different methods, practitioners can enhance confidence in the interpretability and trustworthiness of AI-driven decision-making processes in radiology and other medical domains. (7-9)

III FUTURE XAI POTENTIAL

Evaluation of XAI techniques in radiology represents a crucial step in ensuring their efficacy and reliability. While existing evaluation methods from computer vision offer insights, they may not fully translate to the complexities of radiological tasks. To address this gap, the introduction of "Clinical XAI Guidelines" aims to evaluate XAI techniques in medical imaging based on five key criteria: (1) understandability, (2) clinical relevance, (3) truthfulness, (4) informative plausibility, and (5) computational efficiency. (10) In a recent study, these five criteria were applied to assess sixteen commonly used visual explanation techniques in radiological tasks. However, none of the techniques met all five criteria, underscoring the challenges inherent in developing XAI methods tailored to the specific demands of radiology. (11) This underscores the importance of embracing explainable-by-design approaches, which embed explain ability into AI models from their inception. By integrating explain ability into the development stages of AI models, practitioners can foster transparency, interpretability, and trustworthiness throughout the entire lifecycle of AI-driven decision-making processes in radiology and other medical domains. (1)

Contrary to the notion of an inherent tradeoff between performance and explain ability in AI, recent developments have demonstrated the potential for explainable artificial intelligence (XAI) techniques to enhance AI performance. One compelling approach involves leveraging XAI to improve AI performance. For instance, visual explanations can play a pivotal role in ranking radiological images for active learning, thereby guiding the selection of images that contribute most effectively to model refinement. Similarly, in scenarios involving human-in-the-loop settings with numerous unlabeled images, visual explanations can aid in prioritizing which images to label next, thereby optimizing the annotation process.

Another noteworthy example involves utilizing visual explanations to enforce differential interpretability between classes within each sample. This not only enhances performance but also ensures that the visual explanations align more closely with expert annotations, thereby bolstering the overall accuracy and reliability of the AI model. These innovative applications highlight the transformative potential of XAI in augmenting AI performance while simultaneously enhancing transparency, interpretability, and alignment with domain expertise. By embracing such approaches, practitioners can harness the synergies between performance and explain ability, paving the way for more robust and trustworthy AI systems across various domains, including radiology and medical imaging. (12-14)

Expanding the scope of explainable artificial intelligence (XAI) to include biological explanations represents a promising frontier in medical research and imaging analysis. For instance, pathway analyses of gene expression data obtained from RNA sequencing have unveiled intriguing correlations between MRI characteristics of breast cancer and underlying biological processes. Specifically, features such as contrast enhancement, smoothness, and sharpness of cancer lesions have been linked to ribosome and peptide chain elongation pathways.

This insightful discovery underscores the potential for integrating biological insights into XAI frameworks, enabling a deeper understanding of the intricate relationships between imaging findings and underlying physiological mechanisms. By elucidating the molecular pathways and biological processes driving observable imaging features, XAI enhanced with biological explanations can offer invaluable insights into disease mechanisms, prognosis, and treatment response.

Furthermore, leveraging biological explanations in XAI not only enhances the interpretability and clinical relevance of AI-driven analyses but also fosters interdisciplinary collaboration between radiology, genomics, and molecular biology. Through such collaborative efforts, researchers and clinicians can unlock new avenues for personalized medicine, early detection, and targeted therapies, ultimately improving patient outcomes in oncology and other medical disciplines. (15)

To advance beyond mere correlation and offer explanations that unveil cause-and-effect relationships, the integration of causality in explainable artificial intelligence (XAI) holds tremendous potential. By incorporating causal relationships, radiologists and clinicians can attain a deeper comprehension of the underlying mechanisms driving AI-driven decisions, paving the way for enhanced diagnostic accuracy and clinical decision-making.

One notable advantage of integrating causality into XAI is the capability to uncover and mitigate potential biases inherent in AI algorithms. By discerning causal relationships, practitioners can identify and address biases that may distort or influence model predictions, thereby fostering more equitable and reliable diagnostic outcomes.

One illustrative approach to incorporating causality in XAI involves employing counterfactual explanations. Consider a chest X-ray indicating pleural effusion. A counterfactual explanation would explore how the same chest X-ray image could be altered to prevent the classifier from predicting pleural effusion. This personalized and interactive approach to explanation offers insights into the specific features and characteristics influencing model predictions, empowering radiologists and clinicians to make informed decisions.

By embracing causality in XAI, the medical community can unlock new frontiers in diagnostic interpretation, treatment planning, and patient care, ultimately advancing the capabilities and impact of AI-driven radiology and medical imaging. (16-18)

In summary, explainable artificial intelligence (XAI) represents a dynamic and burgeoning field with immense potential in radiology. As a community, active engagement and collaboration are pivotal to shaping the trajectory of XAI development within the radiology domain. By collectively defining criteria and priorities, we can guide the evolution of XAI techniques to address the unique challenges and requirements of radiological practice.

This collaborative approach ensures that XAI methodologies are tailored to the specific needs of clinicians, radiologists, and patients while adhering to regulatory standards and ethical considerations. By fostering dialogue and consensus-building, we can foster the creation of personalized XAI solutions that enhance diagnostic accuracy, improve patient outcomes, and augment clinical decision-making processes. Ultimately, by championing active participation and collective decision-making, we can harness the full potential of XAI to revolutionize radiology practice, advancing patient care and driving innovation in medical imaging and diagnostics. (19).

REFERENCES

[1] Rudin C (2019) Stop explaining black box machine learning models for high stakes decisions and use interpretable models instead. Nat Mach Intell 1:206–215. https://doi.org/10.1038/s42256-019-0048-x

[2] Litjens G, Kooi T, Bejnordi BE et al (2017) A survey on deep learning in medical image analysis. Med Image Anal 42:60–88. https://doi.org/10.1016/j.media.2017.07.005

[3] Adadi A, Berrada M (2018) Peeking inside the black-box: a survey on explainable artificial intelligence (XAI). IEEE Access 6:52138–52160. https://doi.org/10.1109/ACCESS.2018.2870052

[4] ACM FAccT. https://facctconference.org/. Accessed 12 Jul 2023

[5] Gunning D, Aha DW (2019) DARPA's explainable artificial intelligence (XAI) program. AI Mag 40:44–58. https://doi.org/10.1609/AIMAG.V40I2.2850

[6] Reyes M, Henriques Abreu P, Cardoso J (2022) Interpretability of machine intelligence in medical image computing. 13611:. https://doi.org/10.1007/978-3-031-17976-1

[7] van der Velden BHM, Kuijf HJ, Gilhuijs KGA, Viergever MA (2022) Explainable artificial intelligence (XAI) in deep learning-based medical image analysis. Med Image Anal 79:102470. https://doi.org/10.1016/J.MEDIA.2022.102470

[8] Arun N, Gaw N, Singh P, et al (2021) Assessing the (un) trustworthiness of saliency maps for localizing abnormalities in medical imaging. Radiol Artif Intell e200267

[9] Adebayo J, Gilmer J, Muelly M, et al (2018) Sanity checks for saliency maps. arXiv:1810.03292

[10] Doshi-Velez F, Kim B (2017) Towards a rigorous science of interpretable machine learning. arXiv:1702.08608

[11] Jin W, Li X, Fatehi M, Hamarneh G (2023) Guidelines and evaluation of clinical explainable AI in medical image analysis. Med Image Anal 84:102684. https://doi.org/10.1016/J.MEDIA.2022.102684

[12] Weber L, Lapuschkin S, Binder A, Samek W (2023) Beyond explaining: opportunities and challenges of XAI-based model improvement. Inf Fusion 92:154–176. https://doi.org/10.1016/J.INFFUS.2022.11.013

[13] Mahapatra D, Poellinger A, Shao L, Reyes M (2021) Interpretability-driven sample selection using self supervised learning for disease classification and segmentation. IEEE Trans Med Imaging 40:2548–2562. https://doi.org/10.1109/TMI.2021.3061724

[14] Mahapatra D, Poellinger A, Reyes M (2022) Interpretability-guided inductive bias for deep learning based medical image. Med Image Anal 81:102551. https://doi.org/10.1016/J.MEDIA.2022.102551

[15] Bismeijer T, Van Der Velden BHM, Canisius S et al (2020) Radiogenomic analysis of breast cancer by linking MRI phenotypes with tumor gene expression. Radiology 296:277–287. https://doi.org/10.1148/radiol.2020191453

[16] Chattopadhyay A, Manupriya P, Sarkar A, Balasubramanian VN (2019) Neural network attributions: a causal perspective. arXiv:1902.02302

[17] van Amsterdam WAC, Verhoeff JJC, de Jong PA, Leiner T, Eijkemans MJC (2019) Eliminating biasing signals in lung cancer images for prognosis predictions with deep learning. NPJ Digit Medicine 1(2):1–6. https://doi.org/10.1038/s41746-019-0194-x

[18] Singla S, Wallace S, Triantafillou S, Batmanghelich K (2021) Using causal analysis for conceptual deep learning explanation. Med Image Comput Comput Assist Interv 12903:519. https://doi.org/10.1007/978-3-030-87199-4_49

[19] Gyevnar B, Ferguson N, Schafer B (2023) Get your act together: a comparative view on transparency in the AI act and technology