European Radiology
https://doi.org/10.1007/s00330-023-10121-4

European Radiology
EUROPEAN SOCIETY OF RADIOLOGY

COMMENTARY

Explainable AI: current status and future potential

Bas H. M. van der Velden[1]

Received: 27 June 2023 / Revised: 20 July 2023 / Accepted: 1 August 2023

There has been an increasing trend of using artificial intelligence (AI) in high-stakes decision-making that has an impact on human lives, including but not limited to the criminal justice system, autonomous vehicles, food safety, and radiology [1]. The current standard for AI in radiology is deep learning [2]. Deep learning uses neural networks with many interconnected layers that involve nonlinear relationships. Even if we try to understand and describe these layers and connections, it is unfeasible to fully grasp how the neural network makes its decisions. This is why deep learning is often called a "black box." People are worried that these black boxes might have biases that go unnoticed, which could have serious consequences in high-stakes decision-making [1].

There is a growing demand for methods to improve our understanding of the black box nature of deep learning. These methods are often referred to as explainable artificial intelligence (XAI) [3]. Some notable XAI initiatives include those 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) [4, 5]. For medical imaging, there is a dedicated annual workshop on Interpretability of Machine Intelligence in Medical Image Computing (iMIMIC) at the International Conference on Medical Image Computing and Computer Assisted Intervention (MICCAI) [6].

Current XAI status

Current XAI techniques in radiology typically either provide a visual explanation, a textual explanation, an example-based explanation, or a combination of these [7]. Visual explanations often provide a "heatmap" or "saliency map," pinpointing where the algorithm based its decision on. Visual explanations are currently by far the most used XAI technique in radiology [7]. Textual explanations provide textual descriptions, ranging from relatively simple descriptions such as "hyperintense lesion" up to entire medical reports. Example-based explanations provide relevant examples to explain how a neural network made a decision. It is similar to how a radiologist leverages past cases to analyze the case at hand.

Many XAI methods are post hoc, which means that they provide explanations after a neural network has already been trained. This has several advantages [7]. For example, post hoc XAI techniques are often open source and relatively "plug and play," especially in frameworks such as captum.ai. Furthermore, post hoc XAI is often model agnostic, meaning that it will generate an explanation regardless of the algorithm it is explaining. Therefore, it is possible to provide explanations to neural networks that are currently operational in your clinic or department. There are also some notable disadvantages to post hoc XAI. Post hoc XAI can demonstrate unexpected behavior. For example, not all post hoc XAI techniques demonstrate high validity [8], defined as whether the explanation is correct and corresponds to what the end user expects [7]. Furthermore, there are concerns about robustness [9]. A practical advice to overcome these disadvantages is to examine multiple post hoc XAI techniques and assess the consistency between the explanations.

Future XAI potential

An important step is to evaluate how well an XAI technique performs. Several evaluation methods exist from computer vision [10], but these do not fully translate to radiology. Therefore, "Clinical XAI Guidelines" have recently been proposed [11] to evaluate XAI techniques in medical images based on five criteria: (1) understandability, (2) clinical relevance, (3) truthfulness, (4) informative plausibility, and (5) computational efficiency. These

✉ Bas H. M. van der Velden
bas.vandervelden@wur.nl

[1] Wageningen Food Safety Research, Akkermaalsbos 2, 6708 WB Wageningen, The Netherlands

Published online: 17 August 2023

five criteria were evaluated in radiological tasks for sixteen commonly used visual explanation techniques; none of them met all five criteria [11]. This further reinforces the need for adopting explainable-by-design methods [1], which integrate explainability into AI models from their initial development stages [1].

It is often said that there is an inherent tradeoff between performance and explainability that cannot be avoided. This is not necessarily true: An exciting development is to utilize XAI to improve AI performance [12]. As an example, visual explanations can be used to rank which radiological images should be used next in active learning, leading to a better-performing AI model [13]. This ranking could also be used to select which image to label next, in case of a human-in-the-loop setting with many unlabeled images. Another example uses visual explanations to enforce differentially between visual explanations per class in each sample. This yields better performance, and the visual explanations align more with expert annotations [14].

XAI can be expanded to incorporate biological explanations. As an example, pathway analyses of gene expression data from RNA sequencing revealed that MRI characteristics of breast cancer, such as the contrast enhancement, the smoothness, and the sharpness of the cancer, can be explained by ribosome and peptide chain elongation pathways [15]. This shows the potential of biological processes to be used as explanations.

To go beyond mere correlation and provide explanations that demonstrate cause-and-effect relationships, XAI needs to incorporate causal relationships [16]. By integrating causality in XAI, radiologists can gain a deeper understanding of the underlying mechanisms behind AI-driven decisions. An advantage of incorporating causality is the ability to gain insights into potential biases or to remove such biases [17]. Initial examples of causality in XAI include those using a counterfactual explanation. Let us imagine a chest X-ray showing pleural effusion. How would the same chest X-ray need to appear for the classifier to not predict pleural effusion? This is a counterfactual explanation. Such a counterfactual can provide a personalized and interactive explanation [18].

In summary, explainable artificial intelligence (XAI) is a young, rapidly evolving, and exciting field. It is essential for us as a community to actively contribute to the direction of XAI in the field of radiology. By deciding together on the criteria and aspects that should be prioritized, we can shape the future development of XAI techniques in radiology. This involvement ensures that the emphasis is placed on the specific needs and challenges of the radiology domain, enabling us to create personalized XAI that aligns with the need of clinicians, radiologists, and patients, while complying with regulatory standards [19].

Funding Funding for this research has been provided by the European Union's Horizon Europe research and innovation programme EFRA (Grant Agreement Number 101093026), funded by the European Union. Views and opinions expressed are, however, those of the author only and do not necessarily reflect those of the European Union or Research Executive Agency. Neither the European Union nor the granting authority can be held responsible for them.

Declarations

Guarantor The scientific guarantor of this publication is Bas van der Velden.

Conflict of interest The author of this manuscript declares no relationships with any companies, whose products or services may be related to the subject matter of the article.

Statistics and biometry No complex statistical methods were necessary for this paper.

Informed consent Written informed consent was not required for this study.

Ethical approval Institutional Review Board approval was not required.

Study subjects or cohorts overlap Not applicable.

Methodology
- Invited commentary

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