A Human-Centered Explainable AI Framework for Trustworthy Radiological Decision Support Systems

[1]Babita, [2]Dr. Prakash Mathew

[1]Research Scholar, Department of Radiology & Imaging, North East Christian University

[2]Professor, Department of Radiology & Imaging, North East Christian University

Abstract

Artificial intelligence-driven radiological decision support systems have shown impressive diagnostic accuracy, but their use in clinical settings is still quite limited. This is mainly due to issues like transparency, interpretability, and trust. Many of these systems function as black boxes, offering predictions without providing explanations that are meaningful in a clinical context. This article introduces a human-centered explainable artificial intelligence (XAI) framework aimed at boosting transparency, trust, and accountability in radiological decision support.

The framework combines deep learning-based image analysis with post-hoc explainability techniques to create visual and feature-level explanations that resonate with radiologists' diagnostic thought processes. It prioritizes clinician interaction, ethical accountability, and seamless integration into existing workflows, rather than focusing solely on predictive performance. The study brings together experimental insights from the development of explainable models and assesses the framework from technical, clinical, and social angles. The results indicate that making explainability a fundamental design principle can significantly enhance clinician confidence and support better decision-making, ethical compliance, and enables safer human–AI collaboration in radiological practice.

Keywords: Human-Centered AI; Explainable Artificial Intelligence; Radiological Decision Support; Medical Imaging; Clinical Trust; Ethical AI; Healthcare Systems

1. Introduction

The use of artificial intelligence in radiology has taken off, especially with the rise of deep learning techniques that can handle vast amounts of medical imaging data. Convolutional neural networks (CNNs) have set new benchmarks in areas like disease detection, segmentation, and classification across various imaging types. However, even with these advancements, the real-world adoption of AI systems in clinical settings is still quite limited.

One major hurdle is the lack of interpretability that comes with most deep learning models. Making decisions in radiology is a critical process that demands transparency, justification, and accountability. Black-box AI systems complicate established clinical workflows by providing predictions without the explanations that clinicians need to validate or share with patients.

Human-centered artificial intelligence focuses on creating AI systems that enhance human expertise, honor professional judgment, and align with ethical and social values. Explainable AI is crucial in this context, as it allows clinicians to comprehend, assess, and trust the decisions made with AI assistance. This article builds on these ideas to suggest a human-centered explainable AI framework

for supporting radiological decision-making.

2. The Need for a Human-Centered Explainable AI Approach

Even though deep learning has made remarkable strides in analyzing radiological images, the integration of artificial intelligence into clinical practice is still limited due to the unclear nature of most AI models. Deep neural networks, especially convolutional neural networks, are intricate systems with millions of parameters, making it tough to decipher their internal decision-making processes, even for AI specialists. While these models have shown outstanding performance in image classification and disease detection, their lack of transparency poses significant challenges in clinical settings.

Radiological decision-making is all about interpretation and accountability. Radiologists need to back up their diagnostic conclusions with clear imaging features and solid clinical reasoning. However, black-box AI systems, which churn out predictions without any understandable reasoning, clash with this professional expectation and can shake clinicians' confidence in AI-assisted results. Consequently, even the most accurate AI systems might not be fully embraced or could even be dismissed in real-world clinical settings.

On top of that, using non-explainable AI in healthcare brings up ethical and legal issues. When diagnostic mistakes happen, it’s tough to pinpoint who’s responsible if AI systems can’t offer clear explanations for their decisions. This lack of accountability can jeopardize patient safety and erode trust in AI-driven healthcare technologies. Regulatory bodies and healthcare policymakers are increasingly stressing the importance of transparency, auditability, and human oversight in AI medical systems.

A human-centered approach to explainable AI tackles these issues by focusing on interpretability, clinician engagement, and ethical considerations, all while maintaining predictive accuracy. Explainable AI allows radiologists to grasp the reasoning behind specific decisions, evaluate their clinical relevance, and seamlessly incorporate AI support into their diagnostic processes without sacrificing their professional independence. By centering human users in the design of these systems, explainable AI fosters a safer, more reliable, and ethically sound use of AI in radiological decision support

3. Proposed Human-Centered Explainable AI Framework

To tackle the challenges posed by black-box AI systems in radiological decision-making, this study introduces a human-centered explainable AI framework. This framework aims to weave together transparency, clinical usability, and ethical accountability within radiological decision support systems. It focuses on fostering collaboration between AI systems and radiologists, rather than allowing for independent decision-making by the AI.

3.1 Data Acquisition and Model Development Layer

This layer takes charge of managing radiological imaging data, which includes tasks like preprocessing, normalization, and augmentation to ensure the data is of high quality and robust. We primarily use deep learning models, especially convolutional neural networks, for extracting features and making diagnostic predictions. To enhance performance in situations where labeled

medical data is scarce, we might employ transfer learning strategies.

The main goal of this layer is to deliver dependable diagnostic performance while ensuring it aligns well with the explainability mechanisms that follow.

3.2 Explainability and Interpretation Layer

In this layer, we incorporate post-hoc explanation techniques such as Grad-CAM, SHAP, and LIME to produce outputs that are easy to interpret for model predictions. Visual explanations draw attention to the anatomically significant areas that impact diagnostic results, while feature-based explanations shed light on the importance of various input features.

This layer serves as the core component for transforming black-box predictions into transparent and clinically interpretable information that can be evaluated by radiologists.

3.3 Clinician Interaction and Decision Support Layer

We focus on presenting clear and understandable outputs through user-friendly interfaces that enhance clinician engagement. Radiologists have the ability to visualize explanation maps, evaluate confidence levels, and juxtapose AI-generated results with their own diagnostic insights.

This system is crafted to encourage thoughtful evaluation instead of blind trust in AI predictions, which helps to minimize automation bias and uphold the independence of clinicians.

3.4 Ethical and Accountability Layer

The ethical layer ensures traceability, documentation, and auditability of AI-assisted decisions. This layer supports compliance with healthcare regulations and ethical guidelines by maintaining records of model predictions, explanations, and clinician interactions.

By embedding accountability into system design, the framework aligns AI deployment with medico-legal and ethical requirements in clinical practice.

4. Clinical Significance of the Proposed Framework

We discuss the Clinical Significance of the Proposed Framework. The human-centered explainable AI framework we propose holds considerable promise for clinical radiology. By offering interpretable explanations alongside diagnostic predictions, it empowers radiologists to validate AI outputs and seamlessly incorporate them into their clinical decision-making.

Explainable AI becomes especially crucial in complex, borderline, or high-risk scenarios where diagnostic uncertainty looms large. The visual and feature-based explanations provide extra evidence that can either support or question initial interpretations, ultimately boosting diagnostic confidence and lowering error rates.

Explainable AI shines in situations that are complex, borderline, or carry high risks, especially when there's a lot of uncertainty in diagnostics. By offering visual and feature-based explanations, it adds extra layers of evidence that can either back up or question initial interpretations. This not only boosts diagnostic confidence but also helps to lower error rates.

Moreover, this framework positions AI as a helpful second opinion rather than a substitute for human expertise. This teamwork approach builds trust among clinicians and promotes the thoughtful

integration of AI systems into everyday radiological practices.

Additionally, the framework serves an educational purpose, as the explainable outputs can be utilized to train radiology residents and early-career clinicians. They can learn about clinically significant image features and the reasoning behind diagnoses through these illustrative examples.

5. Ethical, Social, and Regulatory Considerations

The use of artificial intelligence in radiology brings up some important ethical, social, and regulatory questions. The framework being proposed tackles these concerns head-on by making explainability and human oversight key design principles. From an ethical standpoint, explainable AI fosters transparency, accountability, and respects patient autonomy. This means that clinicians can clearly explain AI-assisted decisions to their patients, which helps ensure informed consent and encourages shared decision-making.

On a social level, transparent AI systems help build public trust and acceptance of AI in healthcare. When AI is explainable, it eases the worries people might have about decisions made by algorithms, reinforcing the idea that AI is a helpful clinical tool rather than a mysterious authority.

From a regulatory perspective, explainable AI makes it easier to comply with new healthcare AI regulations that focus on auditability, traceability, and risk management. By keeping thorough documentation of AI decisions and their explanations, this framework supports regulatory reviews and ensures medico-legal accountability.

6. Evaluation and Expected Outcomes

The proposed human-centered explainable AI framework is assessed based on its diagnostic reliability, interpretability, and usability in clinical settings. We evaluate model performance using standard radiological metrics like accuracy, sensitivity, specificity, precision, recall, F1-score, and the area under the ROC curve. These metrics help ensure that our explainable models deliver competitive diagnostic performance compared to traditional black-box systems.

To evaluate explainability, we qualitatively analyze the clinical relevance of visual explanation maps and feature-importance outputs. A key indicator of interpretability is how well the explanation regions align with the anatomical structures identified by radiologists. Moreover, explainability aids in effective error analysis by clarifying whether incorrect predictions stem from ambiguous imaging patterns or irrelevant features.

From a usability standpoint, this framework aims to boost radiologists' trust and confidence by allowing them to critically assess AI predictions. The explainable outputs promote informed decision-making and help minimize the risk of automation bias, fostering safer collaboration between humans and AI in clinical practice.

7. Discussion

The insights and conceptual evaluation of this framework highlight the crucial role of explainability as a core design principle for AI-driven radiological decision support systems. While traditional black-box models focus on performance, they often overlook essential clinical needs like transparency, accountability, and interpretability. Our framework addresses this gap by integrating explainability and

clinician interaction into the system's architecture.

Incorporating explainability does not significantly hinder diagnostic performance, suggesting that transparency and accuracy can coexist. In fact, explainable AI strengthens system robustness by empowering clinicians to identify errors, biases, and limitations that might otherwise go unnoticed.

Additionally, the framework encourages a shift from automation-focused AI to one that emphasizes collaboration, keeping human expertise at the forefront. This transition is crucial for ensuring ethical standards, gaining regulatory approval, and achieving the long-term sustainability of AI systems in radiology.

8. Conclusion and Future Directions

This article introduces a human-centered explainable AI framework designed to enhance transparency, trust, and accountability in radiological decision support systems. By merging deep learning with explainability mechanisms that matter in clinical settings, the framework fosters effective collaboration between humans and AI while maintaining diagnostic accuracy.

Looking ahead, future research should prioritize large-scale clinical validation of the framework across various imaging techniques and healthcare settings. Developing adaptive explanation mechanisms that cater to user expertise and specific clinical situations is another exciting avenue for exploration. Moreover, delving deeper into inherently interpretable AI models could lessen the dependence on post-hoc explanation methods. the proposed framework plays a vital role in creating trustworthy, ethical, and clinically acceptable AI systems in radiology, paving the way for the responsible integration of artificial intelligence into healthcare practices.

9. Limitations of the Study

While the proposed human-centered explainable AI framework is conceptually robust and practically relevant, it does come with certain limitations that need to be recognized. Firstly, the framework has mainly been assessed at a conceptual and experimental level. Conducting large-scale prospective clinical validation in real hospital settings was not within the scope of this study. Consequently, challenges related to workflow integration, time efficiency, and the long-term clinical impact still require further exploration.

The explainability techniques we use are mostly post-hoc. While these methods can shed some light on how models behave, they often fall short of revealing the deeper reasoning processes behind deep learning models. The accuracy of these explanations can fluctuate based on factors like data quality, the architecture of the model, and the specific clinical context.

Now, moving on to the third point, the framework doesn't really tackle the computational challenges that come with providing real-time explanations, especially in healthcare settings where resources are tight. Creating explanations for high-resolution medical images can be quite demanding on computational resources, which could pose scalability issues when it comes to practical applications.

10. Implications for Research and Practice

The proposed framework carries significant implications for both future research and clinical practice. For researchers, it emphasizes the importance of shifting from

a focus solely on accuracy to a more comprehensive evaluation that encompasses interpretability, trust, and ethical considerations. This framework lays a solid groundwork for creating and testing explainable AI systems that meet clinical needs.

In terms of clinical practice, the study underscores the value of explainable AI as a supportive tool for decision-making rather than a standalone diagnostic entity. It promotes the responsible integration of AI by maintaining clinician autonomy, fostering informed choices, and building patient trust. Healthcare organizations can utilize such frameworks to steer the procurement, validation, and governance of AI-driven radiological systems.

11. Ethical Approval, Funding, and Conflict of Interest Statement

This study did not involve any direct patient interactions or prospective clinical trials, so formal ethical committee approval was not necessary. All datasets mentioned were sourced from publicly available or previously approved materials, adhering to ethical guidelines.

The authors confirm that no external funding was received for this research.

Additionally, the authors state that there are no conflicts of interest concerning the publication of this article.

References

1. Litjens, G., Kooi, T., Bejnordi, B. E., et al. (2017). A survey on deep learning in medical image analysis. Medical Image Analysis, 42, 60–88.
2. LeCun, Y., Bengio, Y., & Hinton, G. (2015). Deep learning. Nature, 521(7553), 436–444.
3. Esteva, A., Robicquet, A., Ramsundar, B., et al. (2019). A guide to deep learning in healthcare. Nature Medicine, 25(1), 24–29.
4. Topol, E. (2019). High-performance medicine: The convergence of human and artificial intelligence. Nature Medicine, 25, 44–56.
5. Doshi-Velez, F., & Kim, B. (2017). Towards a rigorous science of interpretable machine learning. arXiv preprint arXiv:1702.08608.
6. Holzinger, A., Langs, G., Denk, H., et al. (2019). What do we need to build explainable AI systems for the medical domain? arXiv preprint arXiv:1712.09923.
7. Selvaraju, R. R., Cogswell, M., Das, A., et al. (2017). Grad-CAM: Visual explanations from deep networks via gradient-based localization. Proceedings of the IEEE International Conference on Computer Vision (ICCV), 618–626.
8. Ribeiro, M. T., Singh, S., & Guestrin, C. (2016). "Why should I trust you?" Explaining the predictions of any classifier. Proceedings of the ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, 1135–1144.
9. Lundberg, S. M., & Lee, S. I. (2017). A unified approach to interpreting model predictions. Advances in Neural Information Processing Systems (NeurIPS), 30.
10. Samek, W., Montavon, G., Vedaldi, A., Hansen, L. K., & Müller, K. R. (2021). Explainable AI: Interpreting, explaining and visualizing deep learning. Springer.

11. Tjoa, E., & Guan, C. (2020). A survey on explainable artificial intelligence (XAI): Toward medical XAI. IEEE Transactions on Neural Networks and Learning Systems, 32(11), 4793–4813.
12. 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.
13. Rajpurkar, P., Irvin, J., Zhu, K., et al. (2017). CheXNet: Radiologist-level pneumonia detection on chest X-rays with deep learning. arXiv preprint arXiv:1711.05225.
14. Shen, D., Wu, G., & Suk, H. I. (2017). Deep learning in medical image analysis. Annual Review of Biomedical Engineering, 19, 221–248.
15. European Commission. (2021). Ethics guidelines for trustworthy artificial intelligence. Brussels.
16. Kelly, C. J., Karthikesalingam, A., Suleyman, M., Corrado, G., & King, D. (2019). Key challenges for delivering clinical impact with artificial intelligence. BMC Medicine, 17, 195.