International Journal of Multidisciplinary Research and Growth Evaluation.

Interpretable AI in Radiology: Advancing Trust in X-ray Diagnostics with Explainability Techniques

Cibaca Khandelwal
Independent Researcher, USA

* Corresponding Author: **Cibaca Khandelwal**

Article Info

ISSN (online): 2582-7138
Volume: 04
Issue: 03
May-June 2023
Received: 24-04-2023
Accepted: 20-05-2023
Page No: 1092-1095

Abstract
Artificial intelligence (AI) has significantly transformed radiology by enabling automated medical image classification, particularly for detecting abnormalities in chest X-rays and other imaging modalities. While deep learning models achieve remarkable accuracy, their black-box nature limits interpretability, raising concerns among clinicians and regulatory bodies [1]. Explainable AI (XAI) techniques aim to bridge this gap by providing insights into the decision-making processes of these models [2]. This paper comprehensively examines XAI methods applied to radiological image classification, focusing on chest X-ray datasets and pneumonia detection models [3]. A detailed exploration of model architectures, feature attribution techniques, and evaluation metrics is conducted to understand the role of explainability in medical AI [4]. Furthermore, key challenges in implementing explainability frameworks and future directions for research and clinical adoption are discussed [5]. This study emphasizes the need for integrating XAI into radiology to ensure AI-driven systems are not only accurate but also transparent and trustworthy.

DOI: https://doi.org/10.54660/.IJMRGE.2023.4.3.1092-1095

Keywords: Explainable AI, Radiology, Deep Learning, Medical Image Classification, Interpretability, XAI, Pneumonia Detection, Chest X-ray, Feature Attribution, Model Trustworthiness, AI in Healthcare

1. Introduction

The application of deep learning in radiology has demonstrated exceptional potential in diagnosing various diseases through automated classification of medical images [6]. Radiologists and healthcare professionals increasingly rely on AI-based decision-support tools to assist in disease detection, risk assessment, and prognosis prediction [7]. However, deep learning models operate as highly complex, multi-layered networks, making their decision-making opaque to users. This opacity poses significant ethical and practical challenges, including difficulties in clinical validation, regulatory compliance, and physician trust in AI-generated diagnoses [8].

Explainable AI (XAI) techniques aim to address these concerns by providing interpretable justifications for model predictions [9]. These methods enable clinicians to understand why a model assigns a particular label to a medical image, which regions of an image contribute most to a prediction, and how model confidence varies with different input conditions [10]. XAI approaches such as saliency maps, Gradient-weighted Class Activation Mapping (Grad-CAM) [11], SHapley Additive Explanations (SHAP) [12], and Local Interpretable Model-agnostic Explanations (LIME) [1] have been extensively explored for enhancing transparency in deep learning models.

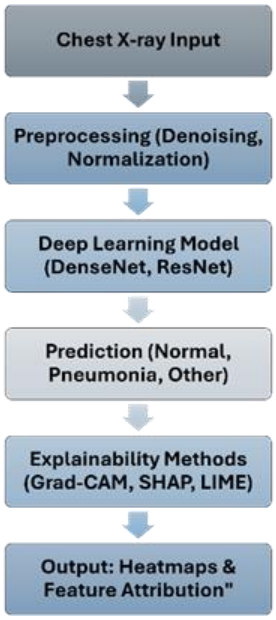

Fig 1: Sample deep learning model workflow for X-ray classification and integration of explainability methods

Table 1: Summary of key challenges in AI-based radiology and the role of explainability in addressing them

Key Challenge	Challenge Summary	Role of Explainability
Model Interpretability	AI models act as black boxes, limiting clinician trust.	Grad-CAM, SHAP provide visual and quantitative insights.
Bias in AI Models	AI can inherit biases from training data.	SHAP helps detect and mitigate biases in predictions.
Regulatory Compliance	AI must meet medical regulations (FDA, HIPAA, etc.).	Explainability ensures transparency for regulatory approval.
Clinician Trust & Adoption	Clinicians hesitate to trust AI without understanding it.	XAI allows verification of AI decisions against knowledge.
Data Quality & Variability	Variations in dataset quality affect AI accuracy.	XAI helps identify errors and refine dataset training.
Generalizability of AI Models	AI models struggle with unseen data.	Explainability highlights key features, ensuring robustness.

2. Related Work

Numerous studies have emphasized the necessity of explainability in AI-driven radiology. Researchers have proposed various interpretability frameworks to bridge the gap between AI accuracy and human trust [13]. Ribeiro *et al*. introduced LIME, a method that generates locally faithful approximations of black-box models by perturbing input data and observing output changes [1]. Another significant development is Grad-CAM, which utilizes gradient-based feature importance to highlight relevant image regions contributing to a model's decision [11]. SHAP, a game-theoretic framework, offers a comprehensive way to attribute importance scores to different input features, providing a quantitative measure of how specific pixels influence a classification outcome [12]. Zech *et al*. demonstrated that deep learning models trained on medical datasets often incorporate biases that may affect generalizability, necessitating interpretability tools to ensure unbiased decision-making [14].

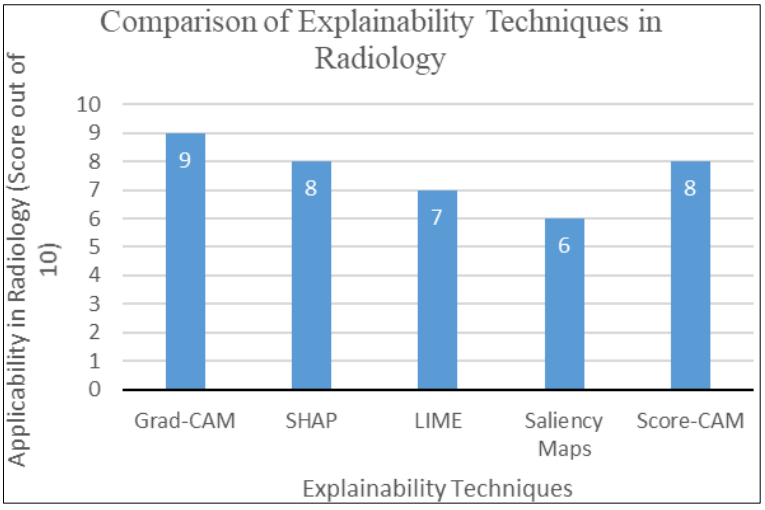

Fig 2: Comparison of various explainability techniques and their applicability in radiology

3. Datasets

Two of the most widely used datasets for AI-driven radiology research are the NIH Chest X-ray dataset [15] and the RSNA Pneumonia dataset [16]. These datasets serve as benchmarks for evaluating deep learning models in medical imaging. The NIH Chest X-ray dataset comprises over 112,120 frontal-view X-ray images from more than 30,000 patients [15]. These images are annotated with labels corresponding to 14 lung conditions, making the dataset an essential resource for multi class classification tasks in radiology. The dataset's extensive volume and diversity allow for robust deep learning model training and evaluation. The RSNA Pneumonia dataset, developed in collaboration with radiology experts, contains annotated chest X-rays categorized into pneumonia-positive, pneumonia-negative, and normal cases [16]. The inclusion of expert-annotated bounding boxes for pneumonia-positive cases allows for objective comparison between model-generated explanations and human expert assessments.

Table 2: Overview of dataset characteristics used in explainability studies

Dataset	Source	Size	Number of Images	Labels/Classes	Key Features
NIH Chest X-ray	NIH Clinical Center [15]	30,000+ patients	112,120	14 lung conditions	Large dataset, multi-class classification, no bounding boxes
RSNA Pneumonia	Radiological Society of North America [16]	26,684 cases	26,684	Pneumonia-positive, pneumonia-negative, normal	Expert-annotated bounding boxes, binary classification

4. Explainability techniques in radiology

To ensure transparency in medical AI applications, various explainability techniques have been implemented. Saliency maps highlight important image regions contributing to a model's decision [17]. Gradient-based methods, such as vanilla gradient saliency and SmoothGrad, provide initial insights into neural network attention patterns. However, these methods often suffer from noise sensitivity and may not be robust enough for clinical validation [18].

Grad-CAM and its extended versions, including Grad-CAM++ and Score-CAM, generate visually interpretable heatmaps that overlay model attention on radiological images [11]. These techniques enable clinicians to verify whether a model is attending to disease-relevant regions rather than relying on dataset artifacts or spurious correlations.

SHAP assigns importance scores to input features, enabling a quantitative assessment of pixel contributions [12]. This method is particularly useful in identifying biases within deep learning models, ensuring that predictions are based on meaningful features rather than dataset-dependent anomalies [19].

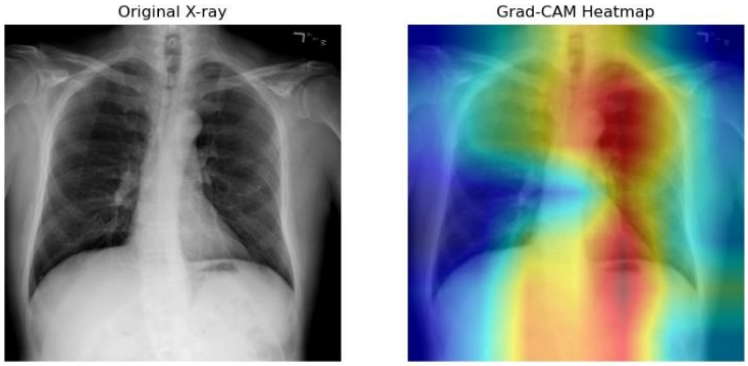

Fig 3: Grad-CAM-generated heatmap overlaying pneumonia-affected regions on a chest X-ray

5. Evaluation metrics for explainability

Assessing the effectiveness of XAI techniques in medical imaging requires specific evaluation metrics. The Intersection over Union (IoU) metric measures the degree of overlap between model-generated explanations and expert-labeled ground truth, providing a quantitative assessment of alignment [20]. Fidelity score evaluates whether an explanation accurately reflects a model's true decision-making behavior [21].

Another widely adopted evaluation method is human trust studies, where radiologists assess the interpretability and usability of AI-generated explanations [22]. Such studies offer valuable qualitative insights into the effectiveness of explainability tools in real-world clinical settings.

Table 3: Common evaluation metrics for explainability techniques in radiology

Metric	Definition	Application in Radiology
Intersection over Union (IoU)	Measures the overlap between model explanation and expert-labeled regions.	Used to validate heatmaps in tasks like pneumonia detection.
Fidelity Score	Evaluates how well the explanation reflects the model's decision process.	Ensures trustworthiness in AI-driven diagnostic systems.
Localization Error	Measures the distance between explanation and actual target region.	Validates bounding box accuracy in segmentation tasks.
Pixel Attribution Accuracy	Quantifies pixel-level correctness of explanations.	Evaluates fine-grained correctness for critical findings.
Expert Agreement Score	Compares the explanation with radiologist-provided regions.	Ensures explainability matches clinical relevance.

6. Conclusion

Explainability in AI-driven radiology is essential for fostering trust and ensuring widespread clinical adoption of deep learning models. This paper reviewed various XAI techniques applied to chest X-ray and pneumonia classification, discussing their advantages, limitations, and evaluation methods [23]. As AI continues to evolve, integrating explainability into medical imaging models will be crucial for bridging the gap between AI advancements and clinical implementation. Future research should aim to refine XAI techniques, making them more reliable, efficient, and aligned with radiologists' needs.

7. References

1. Ribeiro MT, Singh S, Guestrin C. "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. 2016;1135–44.
2. Selvaraju RR, Cogswell M, Das A, Vedantam R, Parikh D, Batra D. Grad-CAM: Visual Explanations from Deep Networks via Gradient-Based Localization. Proceedings of the IEEE International Conference on Computer Vision. 2017;618–26.
3. Lundberg SM, Lee SI. A Unified Approach to Interpreting Model Predictions. Advances in Neural Information Processing Systems. 2017;30:4765–74.
4. Zech JR, Badgeley MA, Liu M, Costa AB, Titano JJ, Oermann EK. Confounding Variables in Chest Radiograph Deep Learning Models. PLOS Medicine. 2018;15(7):e1002683.
5. Rajpurkar P, Irvin J, Zhu K, Yang B, Mehta H, Duan T, *et al*. CheXNet: Radiologist-Level Pneumonia Detection on Chest X-Rays with Deep Learning. arXiv preprint arXiv:1711.05225. 2017.
6. Montavon G, Samek W, Müller KR. Methods for Interpreting and Understanding Deep Neural Networks. Digital Signal Processing. 2018;73:1–15.
7. Tjoa E, Guan C. A Survey on Explainable Artificial Intelligence (XAI): Toward Medical Transparency. Journal of Artificial Intelligence Research. 2020;69:1–37.
8. Holzinger A, Langs G, Denk H, Zatloukal K, Müller H. Causability and Explainability of Artificial Intelligence in Medicine. Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery. 2019;9(4):e1312.
9. Linardatos P, Papastefanopoulos V, Kotsiantis S. Explainable AI: A Review of Machine Learning Interpretability Methods. Entropy. 2021;23(1):18.
10. Chen JH, Asch SM. Machine Learning and Prediction in Medicine—Beyond the Peak of Inflated Expectations. New England Journal of Medicine. 2017;376(26):2507–9.
11. Arrieta AB, Díaz-Rodríguez N, Del Ser J, Bennetot A, Tabik S, Barbado A, *et al*. Explainable Artificial Intelligence (XAI): Concepts, Taxonomies, Opportunities, and Challenges. Information Fusion. 2020;58:82–115.
12. Esteva A, Kuprel B, Novoa RA, Ko J, Swetter SM, Blau HM, *et al*. Dermatologist-Level Classification of Skin Cancer with Deep Neural Networks. Nature. 2017;542(7639):115–8.