International Journal of Information Technology & Computer Engineering

ISSN 2347–3657

Volume 6, Issue 3, 2018

Explainable AI in Medical Diagnostics: A Comparative Evaluation of Interpretability Methods

Aviel D. Rubin

Johns Hopkins University, Baltimore, MD

ABSTRACT

As deep learning models gain traction in medical diagnostics, concerns have arisen regarding their interpretability and trustworthiness—especially in high-stakes domains like radiology and pathology. This paper provides a comparative evaluation of leading explainability techniques applied to convolutional neural networks (CNNs) used in medical image classification. We implement and analyze three popular interpretability methods—Gradient-weighted Class Activation Mapping (Grad-CAM), Layer-wise Relevance Propagation (LRP), and SHAP (SHapley Additive exPlanations)—on CNN models trained on chest X-rays from the NIH ChestX-ray14 dataset to detect conditions such as pneumonia and cardiomegaly. Each method is evaluated based on fidelity, localization accuracy (compared to radiologist-annotated regions), and clinician interpretability via a structured survey with 12 medical professionals. Grad-CAM exhibits the best visual coherence with pathological regions, while LRP provides more granular relevance maps. SHAP offers superior feature-level insights, especially when used with auxiliary patient metadata. However, the added complexity of SHAP explanations made them harder for clinicians to interpret without training. All three methods improve clinician trust compared to black-box outputs alone. The study concludes that no single method is universally superior—each has strengths depending on the task, data modality, and end-user profile. Our findings suggest a hybrid approach, combining region-based and feature-based explanations, may offer the most robust path toward explainable AI in healthcare. This work informs both developers and clinical stakeholders seeking to safely integrate AI into diagnostic workflows.

2. INTRODUCTION

Deep learning has shown remarkable potential in medical diagnostics, particularly in tasks involving medical image classification such as identifying pneumonia in chest X-rays or detecting tumors in histopathology slides. However, the complexity of these models—especially convolutional neural networks (CNNs)—has raised significant concerns about their **interpretability**, **trustworthiness**, and **clinical applicability**. In high-stakes environments like healthcare, it is not sufficient for an algorithm to be accurate; it must also be **explainable** to the clinicians who depend on it for patient care.

Explainable AI (XAI) aims to bridge the gap between black-box models and human understanding. For clinicians to confidently integrate AI tools into diagnostic workflows, it is essential that models provide **transparent, interpretable outputs** that align with clinical reasoning. A growing body of research has proposed methods for generating post hoc explanations from CNN-based classifiers. However, there is little consensus on how these methods compare, particularly in the medical domain where interpretability requirements differ from general computer vision.

This paper evaluates three leading interpretability techniques—**Grad-CAM**, **Layer-wise Relevance Propagation (LRP)**, and **SHAP**—applied to CNNs trained on the NIH ChestX-ray14 dataset. We assess each method's ability to highlight diagnostically relevant features, correlate with expert-annotated disease regions, and support clinician decision-making. By combining empirical evaluation and structured feedback from 12 medical professionals, we aim to provide actionable insights for both AI developers and healthcare practitioners.

3. COMPARISON CRITERIA

To ensure a structured and meaningful comparison across interpretability methods, we define the following evaluation criteria:

1. **Fidelity**
   How accurately does the explanation reflect the true decision-making process of the model? We evaluate this using occlusion sensitivity analysis and logit impact after explanation region masking.
2. **Localization Accuracy**
   How well do the generated heatmaps align with radiologist-annotated pathological regions in the chest X-rays? We use Intersection over Union (IoU) scores with bounding box annotations as the metric.
3. **Clinician Interpretability**
   How understandable and actionable are the explanations to practicing clinicians? This is assessed via a structured survey of 12 medical professionals who rated each method based on clarity, clinical utility, and decision support.
4. **Computational Overhead**
   What is the average time and resource cost of generating explanations during inference? Metrics include explanation latency (ms/image) and GPU/CPU utilization.
5. **Modality Integration**
   Can the method incorporate non-image data (e.g., patient metadata) into its explanations? This factor is crucial in multimodal diagnostic settings.

   These criteria provide a comprehensive framework for evaluating the utility and practicality of explainability methods in real-world clinical AI deployments.

4. METHODOLOGY

4.1 Dataset and Preprocessing

We use the **NIH ChestX-ray14** dataset, a widely adopted benchmark for medical image classification that contains over 100,000 frontal chest X-rays annotated with 14 disease labels. For our experiments:

- Images were resized to **224×224** pixels and normalized.
- The training set comprised 80% of the data, with 10% used for validation and 10% for testing.
- Pathologies evaluated include **pneumonia**, **cardiomegaly**, and **infiltration**.
- Radiologist-annotated bounding boxes (provided in a subset) were used for evaluating localization accuracy.

4.2 Model Architecture

A **ResNet-50 CNN** pretrained on ImageNet was fine-tuned for multi-label classification using binary cross-entropy loss. Training was conducted on an NVIDIA GTX 1080 Ti using PyTorch. The model achieved **AUC scores** of 0.83 (pneumonia), 0.88 (cardiomegaly), and 0.79 (infiltration), consistent with prior work.

4.3 Explanation Methods

We implemented three interpretability techniques:

- **Grad-CAM**: A visual saliency method that produces class-discriminative heatmaps using gradients flowing into convolutional layers.
- **Layer-wise Relevance Propagation (LRP)**: A backward-propagation technique that redistributes model output relevance through the network layers to the input pixels.
- **SHAP**: A unified framework based on Shapley values, approximating each feature's contribution to a given prediction. SHAP was adapted for CNNs via Deep SHAP and used in conjunction with auxiliary patient metadata when available.

4.4 Clinician Survey Design

Twelve board-certified clinicians (radiologists and internists) participated in a structured evaluation. They were presented with **X-ray images and associated explanation maps** generated by each method and asked to score them on:

- Visual clarity (1–5 scale)
- Diagnostic alignment (1–5 scale)
- Likelihood of increasing trust in AI prediction (1–5 scale)

Free-text responses were collected for qualitative feedback.

5. CASE A: GRAD-CAM

Gradient-weighted Class Activation Mapping (Grad-CAM) is one of the most widely used visual explanation methods in computer vision. It leverages the spatial hierarchy of CNNs by computing the gradients of the target class with respect to feature maps in the last convolutional layer. These gradients are globally averaged to produce a coarse heatmap that highlights image regions most influential to the model's decision.

In our experiments, Grad-CAM maps were overlaid on the original X-ray images to create a transparent red heatmap. Results show that Grad-CAM consistently highlights clinically relevant areas in cases of **pneumonia and cardiomegaly**, especially in regions corresponding to lung opacities and enlarged cardiac silhouettes.

Quantitative Evaluation:

- **IoU with radiologist annotations**: 0.42 (pneumonia), 0.47 (cardiomegaly)
- **Fidelity score (logit drop on masked area)**: 21.4%
- **Explanation generation time**: ~15ms/image

Clinician Feedback:

Grad-CAM was rated as the most **visually intuitive** method. Clinicians reported that the coarse heatmaps effectively indicated areas of concern without overwhelming the image. Several noted that while resolution was limited, the transparency overlay allowed them to mentally correlate the highlighted regions with anatomical landmarks.

Grad-CAM's simplicity and visual alignment with radiological intuition make it a strong candidate for real-time clinical decision support—especially where interpretability needs to be immediate and accessible.

6. CASE B: LAYER-WISE RELEVANCE PROPAGATION (LRP)

Layer-wise Relevance Propagation (LRP) redistributes the prediction score backward through the network, assigning relevance values to each input pixel in proportion to its contribution to the output. Unlike gradient-based

methods, LRP operates by decomposing the final prediction score without requiring backpropagation of gradients, producing **high-resolution saliency maps** that offer fine-grained interpretability.

In our implementation, we used the ε-rule variant of LRP to ensure numerical stability and relevance conservation across layers. The resulting saliency maps highlighted detailed anatomical structures such as localized lung infiltrates and enlarged heart boundaries, especially in cases where pathologies overlapped with multiple visual features.

Quantitative Evaluation:

- **IoU with radiologist annotations**: 0.39 (pneumonia), 0.45 (cardiomegaly)
- **Fidelity score (logit drop on masked area)**: 23.7%
- **Explanation generation time**: ~110ms/image

Clinician Feedback:

Clinicians appreciated the **granular detail** provided by LRP, particularly in borderline or ambiguous cases. However, some participants reported that LRP maps were more difficult to interpret visually due to noise and lack of clear spatial boundaries. The lack of smooth gradients occasionally led to confusion about the actual area of clinical concern.

LRP offers superior analytical detail, making it suitable for expert reviews or secondary validation, but may require additional visual post-processing for deployment in clinical interfaces.

7. CASE C: SHAP (SHAPLEY ADDITIVE EXPLANATIONS)

SHAP explains the output of a model by computing the contribution of each input feature using approximations of Shapley values from cooperative game theory. Unlike Grad-CAM or LRP, SHAP is model-agnostic and capable of integrating multimodal data, including image features, patient metadata (e.g., age, gender), and clinical history.

For this study, **Deep SHAP** was implemented using a hybrid of model-specific and sampling-based approximations. We augmented the image-based CNN with auxiliary demographic features and used SHAP to attribute contributions across both modalities.

Quantitative Evaluation:

- **IoU with radiologist annotations** (image component): 0.36 (pneumonia), 0.41 (cardiomegaly)
- **Fidelity score**: 25.5%
- **Explanation generation time**: ~500ms/image (due to sampling and feature attribution)

Clinician Feedback:

Clinicians valued the **contextual depth** offered by SHAP—particularly when image features were combined with patient information. However, many noted that the **complexity of the output** (e.g., force plots, dependency plots) was **difficult to interpret** without prior training. The method was preferred by more technical users but was rated lowest in immediate clinical usability.

SHAP holds promise for **holistic interpretability**, especially in multimodal AI systems, but requires simplification or guided interpretation for real-time clinical use.

8. COMPARATIVE ANALYSIS

A cross-method analysis reveals clear trade-offs among Grad-CAM, LRP, and SHAP when applied to medical diagnostics:

Criterion	Grad-CAM	LRP	SHAP
Visual Clarity	High	Medium	Low
Fidelity	Moderate	High	Highest
Localization (IoU)	0.42 / 0.47	0.39 / 0.45	0.36 / 0.41
Clinician Trust Rating	4.6 / 5.0	4.2 / 5.0	3.4 / 5.0
Explanation Speed	Fast (~15ms)	Moderate (~110ms)	Slow (~500ms)
Metadata Integration	Limited	None	Full

Grad-CAM is best suited for **real-time, visual diagnostics**, especially in tools used directly by radiologists. LRP offers **deep insight** at a pixel level and is valuable in expert-driven diagnostics or second opinions. SHAP stands out for **data-rich interpretability**, ideal for **hybrid systems** combining image and clinical records.

Clinicians favored explanations that were **visually clear and aligned with anatomical landmarks**. While all methods improved trust over black-box outputs, Grad-CAM received the highest average trust scores. SHAP, despite offering deeper insights, was hindered by its complexity and abstract representation style.

9. CONCLUSION

This paper evaluated three leading explainability techniques—Grad-CAM, LRP, and SHAP—applied to CNNs trained for medical image classification. Using a unified experimental design across a benchmark chest X-ray dataset and feedback from domain experts, we assessed each method based on fidelity, localization accuracy, clinician interpretability, and practical deployment concerns.

Our findings confirm that **no single interpretability method is universally optimal**. Each technique excels under specific circumstances:

- **Grad-CAM**: Ideal for intuitive, visual explanations directly aligned with diagnostic imaging workflows.
- **LRP**: Provides detailed, high-resolution relevance maps beneficial in complex or ambiguous cases.
- **SHAP**: Enables comprehensive, multimodal interpretability but requires simplification for clinical use.

We recommend a **hybrid interpretability approach** for medical AI systems—combining Grad-CAM's region-level visualizations with SHAP's contextual insights. This would allow clinicians to quickly assess predictions and drill down into the reasoning behind them when necessary.

For AI developers and clinical stakeholders, the key takeaway is that interpretability must be **task-specific, user-informed, and embedded into clinical context**. Explainable AI is not a one-size-fits-all solution but a design paradigm that must evolve in tandem with clinical practice.

REFERENCES

1. Bach, S., Binder, A., Montavon, G., Klauschen, F., Müller, K. R., & Samek, W. (2015). On pixel-wise explanations for non-linear classifier decisions by layer-wise relevance propagation. *PLOS ONE*, 10(7), e0130140.
2. Boehringer, D., Kather, J. N., & Ghafoorian, M. (2018). Explainable deep learning models in medical imaging. *Proceedings of the Medical Imaging with Deep Learning Conference*.
3. Munnangi, S. (2016). Adaptive case management (ACM) revolution. NeuroQuantology, 14(4), 844–850. https://doi.org/10.48047/nq.2016.14.4.974

4. Ching, T., Himmelstein, D. S., Beaulieu-Jones, B. K., Kalinin, A. A., Do, B. T., Way, G. P., ... & Greene, C. S. (2018). Opportunities and obstacles for deep learning in biology and medicine. *Journal of The Royal Society Interface*, 15(141), 20170387.
5. Doshi-Velez, F., & Kim, B. (2017). Towards a rigorous science of interpretable machine learning. *arXiv preprint arXiv:1702.08608*.
6. Gunning, D. (2017). Explainable artificial intelligence (XAI). *Defense Advanced Research Projects Agency (DARPA), Program Information*.
7. Holzinger, A., Biemann, C., Pattichis, C. S., & Kell, D. B. (2017). What do we need to build explainable AI systems for the medical domain? *arXiv preprint arXiv:1712.09923*.
8. Kermany, D. S., Goldbaum, M., Cai, W., Valentim, C. C. S., Liang, H., Baxter, S. L., ... & Zhang, K. (2018). Identifying medical diagnoses and treatable diseases by image-based deep learning. *Cell*, 172(5), 1122–1131.e9.
9. Lundberg, S. M., & Lee, S. I. (2017). A unified approach to interpreting model predictions. *Advances in Neural Information Processing Systems*, 30, 4765–4774.
10. Montavon, G., Lapuschkin, S., Binder, A., Samek, W., & Müller, K. R. (2017). Explaining nonlinear classification decisions with deep Taylor decomposition. *Pattern Recognition*, 65, 211–222.
11. Rajpurkar, P., Irvin, J., Zhu, K., Yang, B., Mehta, H., Duan, T., ... & Ng, A. Y. (2017). CheXNet: Radiologist-level pneumonia detection on chest X-rays with deep learning. *arXiv preprint arXiv:1711.05225*.
12. Bellamkonda, S. (2018). Data Security: Challenges, Best Practices, and Future Directions. International Journal of Communication Networks and Information Security, 10, 256-259.
13. 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.
14. Samek, W., Wiegand, T., & Müller, K. R. (2017). Explainable artificial intelligence: Understanding, visualizing and interpreting deep learning models. *arXiv preprint arXiv:1708.08296*.
15. 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*, 618–626.
16. Simonyan, K., Vedaldi, A., & Zisserman, A. (2014). Deep inside convolutional networks: Visualising image classification models and saliency maps. *arXiv preprint arXiv:1312.6034*.
17. Zeiler, M. D., & Fergus, R. (2014). Visualizing and understanding convolutional networks. *European Conference on Computer Vision*, 818–833.