*FUDMA Journal of Sciences (FJS)*
*ISSN online: 2616-1370*
*ISSN print: 2645 - 2944*
*Vol. 9 No. 11, November, 2025, pp 461 – 465*
*DOI: https://doi.org/10.33003/fjs-2025-0911-4213*

# EXPLAINABLE AI IN HEALTHCARE: BRIDGING THE GAP BETWEEN MODEL ACCURACY AND INTERPRETABILITY

***[1]Ayanlowo, Emmanuel A. [2]Onalaja, Olawale O. and [3]Obadina, Gabiel O.**

[1]Department of Basic Sciences, Babcock University, Ilishan-Remo, Ogun State. Nigeria.
[2]Department of Computer Science, Ogun State Polytechnic of Health and Allied Science, Ijebu, Ogun State, Nigeria.
[3]Department of Statistics, Olabisi Onabanjo University, Ago-Iwoye, Ogun State, Nigeria.

*Corresponding authors' email: ayanlowoe@babcock.edu.ng

## ABSTRACT

Artificial intelligence (AI) is transforming healthcare by enabling highly accurate diagnostics, personalised treatment planning, and efficient clinical operations. Yet the opacity of advanced machine-learning models remains a barrier to trust and widespread adoption. This paper provides a structured review of explainable AI (XAI) techniques that reconcile predictive strength with interpretability. We examine model-agnostic methods, including Local Interpretable Model-Agnostic Explanations (LIME) and SHapley Additive exPlanations (SHAP), as well as model-specific approaches such as attention mechanisms and Gradient-weighted Class Activation Mapping (Grad-CAM). Empirical evidence drawn from recent clinical studies demonstrates that XAI significantly enhances decision-making. In radiology, Grad-CAM visualisations increased clinician confidence by 30%, while SHAP explanations in electronic health record diagnostics improved trust by 25%. Large-scale chest X-ray experiments (10,000 images) demonstrated that SHAP and LIME achieved high predictive accuracies of 90% and 89%, respectively, compared to 92% for a baseline deep neural network, while providing markedly higher interpretability scores. Patient-centred trials further revealed a 25% improvement in diabetes treatment adherence when AI recommendations were accompanied by high-quality explanations, with compliance rising 5% for every one-point increase in explanation quality ($\beta_1 = 0.05$). These results confirm that XAI can strengthen clinician trust and patient engagement with only minimal loss in accuracy. Remaining challenges include computational cost, absence of standardised interpretability metrics, and evolving regulatory requirements. We recommend the development of hybrid models with intrinsic interpretability, co-designed evaluation frameworks, and educational initiatives to prepare clinicians and patients to act on XAI outputs.

**Keywords**: Explainable AI, SHAP, LIME, Grad-CAM, Clinical Decision Support

## INTRODUCTION

Artificial intelligence (AI) is becoming a cornerstone of modern healthcare, driving innovations that range from precision diagnostics to personalised treatment pathways and more efficient hospital management. Sophisticated machine-learning (ML) algorithms, particularly deep learning architectures, now deliver state-of-the-art performance in tasks such as medical image interpretation, electronic health record (EHR) analytics, and disease risk prediction (LeCun et al., 2015; Huang et al., 2019). Deep convolutional neural networks can detect subtle radiographic abnormalities with accuracy that rivals or even surpasses expert clinicians (Rajpurkar et al., 2017), while recurrent and transformer-based models enable early detection of conditions such as sepsis and cardiac arrhythmias by mining high-dimensional, temporally structured EHR data (Caruana et al., 2015). These breakthroughs promise earlier interventions, improved outcomes, and lower costs across diverse clinical domains.

Yet the very complexity that gives these models their predictive power also renders them difficult to interpret. Deep neural networks often operate as "black boxes", mapping inputs to outputs through layers of nonlinear transformations that defy intuitive understanding (Holzinger et al., 2019). When algorithmic recommendations affect diagnoses, treatment plans, or resource allocation, opacity can undermine trust among clinicians, patients, and regulators. Clinicians are reluctant to rely on decisions they cannot interrogate, while patients may resist AI-driven care if they cannot comprehend how conclusions are reached. This tension is heightened in safety-critical settings such as oncology or intensive care, where incorrect or biased predictions can have immediate and severe consequences (Amann et al., 2020).

Regulatory frameworks increasingly codify the need for transparency. The European Union's General Data Protection Regulation (GDPR) articulates a "right to explanation", obliging organisations to provide meaningful information about the logic behind automated decisions (Goodman & Flaxman, 2017). Similar guidance appears in U.S. Food and Drug Administration (FDA) proposals for adaptive AI-based medical devices and in the World Health Organization's recommendations for trustworthy AI in health. Meeting these requirements demands methods that illuminate the internal reasoning of complex models without sacrificing their predictive accuracy.

Explainable AI (XAI) offers a compelling response. Rather than abandoning high-performing black-box models, XAI techniques generate human-interpretable explanations of their outputs. Model-agnostic approaches such as Local Interpretable Model-Agnostic Explanations (LIME) and SHapley Additive exPlanations (SHAP) provide feature-level attributions applicable across model types (Ribeiro et al., 2016; Lundberg & Lee, 2017). Model-specific strategies, including attention mechanisms and gradient-based saliency methods, embed interpretability directly into neural network architectures (Bahdanau et al., 2014; Selvaraju et al., 2017). Recent empirical studies demonstrate that such methods can increase clinician confidence in AI recommendations (Kim et al., 2023; Lee et al., 2024) and improve patient adherence when explanatory feedback accompanies automated advice (Patel et al., 2024).

This paper situates XAI at the intersection of technological capability and clinical necessity. We survey leading methodologies, examine their mathematical underpinnings, and assess evidence of their performance in real-world healthcare applications. By critically evaluating their strengths, limitations, and regulatory implications, we highlight both the opportunities and the unresolved challenges of embedding explainable AI within routine clinical practice.

**Literature Review**

The emergence of explainable artificial intelligence (XAI) stems from longstanding concerns about the opacity of complex machine-learning (ML) systems and their suitability for high-stakes decision-making. Historically, interpretability in predictive modelling was addressed through intrinsically transparent algorithms, such as linear regression, decision trees, and rule-based systems, that allow straightforward mapping from input variables to outcomes (Molnar, 2020). These models provide clear parameter estimates and human-readable decision rules, making them naturally amenable to clinical audit and regulatory review. However, the dramatic performance gains of deep neural networks in tasks like image classification and sequential data analysis shifted attention toward powerful but opaque models (LeCun et al., 2015; Rajpurkar et al., 2017). As healthcare applications increasingly demand the accuracy of these complex systems, researchers have sought methods to "open the black box" without sacrificing predictive power.

**Foundations of Explainability**

XAI draws on a diverse set of disciplines to produce interpretable explanations. Statistical concepts underpin methods that quantify feature importance or sensitivity, while game theory informs allocation of contributions across feature subsets. Information theory and human–computer interaction also guide the design of explanations that are not merely mathematically rigorous but also cognitively meaningful to end-users such as clinicians (Doshi-Velez & Kim, 2017).

Two landmark contributions remain central. Local Interpretable Model-Agnostic Explanations (LIME) introduced by Ribeiro et al. (2016) approximates the behaviour of a complex model $f(x)$ in the vicinity of an instance xxx with a simpler surrogate $g(x)$, typically a sparse linear model. By perturbing inputs around the point of interest and weighting them according to proximity, LIME provides an intuitive local explanation of how each feature influences the prediction. SHapley Additive exPlanations (SHAP), proposed by Lundberg and Lee (2017), extends cooperative game theory to model interpretation, distributing the output prediction among features based on their marginal contributions across all possible feature coalitions. SHAP values have become a de facto standard for global and local interpretability because of their solid axiomatic foundation and consistency guarantees.

**Model-Specific Approaches**

In parallel, model-specific techniques integrate explanation directly into neural architectures. Attention mechanisms, first described by Bahdanau et al. (2014) for neural machine translation, learn a set of weights highlighting input components most relevant to a prediction. This concept readily transfers to healthcare tasks such as clinical text mining or genomic sequence analysis, where it can reveal clinically significant words or motifs (Huang et al., 2019). Gradient-based visualisation methods such as Grad-CAM (Selvaraju et al., 2017) exploit back-propagated gradients to generate class-discriminative heatmaps over medical images, enabling radiologists to see which regions drive a model's decision. These methods are valued for their ability to provide intuitive, visually anchored explanations without retraining the model.

**Healthcare-Specific Considerations**

Healthcare literature underscores that interpretability is not merely a technical preference but a clinical and ethical imperative. Holzinger et al. (2019) emphasise "causability", the alignment between computational explanations and causal reasoning demanded in medicine, arguing that explanations must be comprehensible to domain experts, not only to data scientists. Amann et al. (2020) further highlight that legal and regulatory frameworks, including the European Union's General Data Protection Regulation (GDPR), enshrine a "right to explanation", obliging developers to provide meaningful insights into algorithmic decisions.

Empirical studies increasingly show that XAI methods can enhance user trust and engagement. Kim et al. (2023) demonstrated that visual explanations via Grad-CAM increased radiologists' confidence in AI-based chest X-ray diagnostics by approximately 30 %. Similarly, Lee et al. (2024) reported a significant improvement in clinicians' acceptance of AI-driven EHR diagnostics when SHAP explanations accompanied model outputs. Beyond clinicians, patient outcomes also benefit: Patel et al. (2024) found that diabetes patients receiving AI-generated recommendations with accompanying explanations exhibited a 25 % increase in treatment adherence, underscoring the broader public-health implications of interpretability.

**Ongoing Debates and Challenges**

Despite these advances, a persistent debate concerns the trade-off between interpretability and predictive performance. Rudin (2019) argues that, for high-stakes domains like healthcare, the use of post-hoc explanations for inherently opaque models is insufficient, advocating instead for intrinsically interpretable models that can achieve competitive accuracy. Others counter that hybrid strategies, combining transparent components with deep networks, may offer a pragmatic path forward (Molnar, 2020). Additional challenges include the lack of universally accepted metrics for explanation quality, the computational burden of methods such as SHAP, and the difficulty of ensuring that explanations are not only technically accurate but also clinically meaningful.

Collectively, the literature reveals a field in rapid evolution: one that balances mathematical rigour, computational feasibility, and the ethical mandate for transparency. As AI systems continue to penetrate clinical workflows, developing robust, standardised, and user-centred explanation methods remains a critical frontier for both research and practice.

**MATERIALS AND METHODS**

This study employs a structured narrative review enriched with formal mathematical exposition to examine explainable artificial intelligence (XAI) techniques in healthcare. The approach integrates conceptual analysis, mathematical formulation, and synthesis of empirical findings from peer-reviewed studies published between 2023 and 2024. Our objective is to clarify the operational principles of prominent XAI methods and to evaluate their clinical impact in terms of (1) improvements in clinician trust, (2) comparative model performance, and (3) patient behavioural outcomes.

**Methodological Framework**

Relevant literature was identified through searches of *PubMed*, *IEEE Xplore*, and *ACM Digital Library* using combinations of the keywords *explainable AI*, *healthcare*, *LIME*, *SHAP*, *attention*, *Grad-CAM*, and *interpretability*. Studies were included if they (i) applied an XAI technique to a healthcare domain, (ii) reported quantitative performance metrics, and (iii) described either clinician trust or patient outcomes. Results were synthesised narratively, with special attention to mathematical definitions of the methods reviewed.

**Taxonomy of XAI Techniques**

XAI approaches were categorised into model-agnostic and model-specific families to distinguish methods that can be applied to any predictive model from those tailored to particular architectures.

**RESULTS AND DISCUSSION**

This section synthesises empirical findings on the clinical impact of explainable artificial intelligence (XAI), focusing on clinician trust, model performance and patient behavioural outcomes. Evidence is drawn from peer-reviewed studies and large experimental datasets published between 2023 and 2024.

**Model-Agnostic Methods**

These methods operate independently of the underlying predictive model $f(x)$, making them broadly applicable across classifiers and regressors.

***Local Interpretable Model-Agnostic Explanations (LIME)***

LIME constructs a simple surrogate model $g(x)$ (e.g., a sparse linear regressor) that locally approximates the complex black-box model $f(x)$ around a target instance $x_0$. Given a set of perturbed samples $Z = \{z_1, \dots, z_m\}$ and a locality kernel $\pi_{x_0}(z)$ that down-weights distant points, LIME minimises a locality-weighted squared loss:

$$\mathcal{L}(f, g, \pi_{x_0}) = \sum_{z \in Z} \pi_{x_0}(z)[f(z) - g(z)]^2$$

subject to a complexity constraint $\Omega(g)$ that encourages interpretability (e.g., sparsity in the coefficient vector). The explanation is the set of non-zero coefficients of $g$, which indicate the locally influential features (Bahdanau et al., 2014).

***SHapley Additive exPlanations (SHAP)***

SHAP applies cooperative game theory to allocate the model output among input features. For a model $f$ with feature set $N$ and a particular feature $i \in N$, the Shapley value $\phi_i$ is:

$$\phi_i = \sum_{S \subseteq N/\{i\}} \frac{|S|!\,(|N| - |S| - 1)!}{|N|!}[f(S \cup \{i\}) - f(S)]$$

where $f(S)$ is the expected model output conditional on the subset SSS of features. This equation ensures three desirable properties, efficiency, symmetry, and additivity, which together yield a fair attribution of the prediction to each feature (Zhang et al., 2024).

**Model-Specific Methods**

These techniques leverage architectural properties of neural networks to derive explanations from within the model itself.

***Attention Mechanisms***

Originally developed for neural machine translation (Bahdanau et al., 2014), attention assigns learnable weights to input tokens, highlighting the elements most relevant to the model's decision. For an input sequence $\{x_1, \dots, x_n\}$ and hidden representations $\{h_1, \dots, h_n\}$, attention scores $\alpha_i$ are computed as

$$e_i = v^\top \tanh(W h_i), \alpha_i = \frac{exp(e_i)}{\sum_{j=1}^{n} exp(e_j)}$$

where $W$ and $v$ are learnable parameters. The context vector $c = \sum_{i=1}^{n} \alpha_i h_i$ represents a weighted summary of salient features. High $\alpha_i$ values point to clinically important words, lab results, or genomic markers.

***Gradient-Weighted Class Activation Mapping (Grad-CAM)***

Grad-CAM generates a class-discriminative heatmap for a convolutional neural network (CNN) by exploiting gradients of the target class score $y^c$ with respect to the final convolutional feature maps $A^k$. The weight for feature map $k$ is:

$$\alpha_k^c = \frac{1}{Z} \sum_i \sum_j \frac{\partial y^c}{\partial A_{ij}^k}$$

where $Z$ is the number of spatial locations. The saliency map is then

$$L_{Grad-CAM}^c = ReLU\left(\sum_k \alpha_k^c A^k\right)$$

which highlights image regions most influential to the class prediction, aiding radiologists in visually validating the model's reasoning.

**Empirical Evidence**

Empirical synthesis focused on clinical studies that evaluated XAI methods in practice. Three key outcome dimensions guided our review:

Clinician Trust: Measured through surveys or behavioural metrics capturing confidence in AI-assisted diagnoses (e.g., Kim et al., 2023).

Model Performance Relative to Baseline: Comparison of accuracy, sensitivity, and computation time of XAI-enhanced models versus standard deep networks, such as in large-scale chest X-ray classification tasks (Zhang et al., 2024).

Patient Behavioural Outcomes: Impact on patient adherence and satisfaction when AI recommendations are accompanied by explanations, exemplified by a 2024 diabetes management trial reporting a 25 % improvement in compliance (Patel et al., 2024).

Key datasets included multi-institutional radiology image repositories (≈10,000 chest X-rays) and anonymised EHR diagnostic trials conducted in North America and East Asia between 2023 and 2024.

**Clinician Trust**

Integrating XAI into diagnostic workflows produced measurable gains in clinician confidence. In a controlled reader study involving 38 radiologists, Grad-CAM visual heatmaps were incorporated into a chest X-ray classification system. When compared with standard outputs, the presence of saliency maps increased mean self-reported confidence in AI-assisted diagnoses by 30 per cent (Kim et al., 2023). A complementary investigation of SHAP-based feature attributions in electronic health record (EHR) diagnostics reported a 25 per cent rise in clinician trust scores on a five-point Likert scale relative to unannotated probability outputs (Lee et al., 2024). These results, summarised in Table 1, demonstrate that transparent case-specific explanations reduce the psychological barrier to relying on automated recommendations and support collaborative clinical decision-making.

**Table 1: Improvement in Clinician Trust when XAI Explanations Accompany AI-driven Decisions**

| Application | XAI Method | Trust Increase (%) |
|---|---|---|
| Radiology (chest X-ray) | Grad-CAM | 30 |
| EHR diagnostics | SHAP | 25 |

**Model Performance**

The trade-off between predictive accuracy, interpretability and computational efficiency was evaluated on a chest X-ray dataset of 10 000 images (Zhang et al., 2024). Accuracy is reported as mean classification performance across five disease categories, interpretability reflects structured clinician feedback, and computation time indicates average per-case explanation overhead. Results appear in Table 2.

**Table 2: Comparative Performance of XAI Methods on Chest X-ray Classification**

| Method | Accuracy (%) | Interpretability | Computation Time (s) |
|---|---|---|---|
| Deep neural network (baseline) | 92 | Low | 0.5 |
| SHAP | 90 | High | 2.0 |
| LIME | 89 | High | 1.8 |
| Grad-CAM | 91 | Medium | 0.8 |

The results show only marginal reductions in accuracy when explainability is introduced. SHAP and LIME maintained high predictive power while providing the greatest interpretability, though both incurred longer computation times, with SHAP averaging two seconds per case. Grad-CAM achieved a balanced profile, maintaining 91 per cent accuracy and moderate interpretability with the lowest computational overhead of 0.8 seconds.

**Patient Outcomes**

Evidence from a 2024 randomised controlled trial of diabetes management indicates that explainability can also improve patient behaviour (Patel et al., 2024). Participants received AI-generated insulin dosage recommendations either with or without explanatory feedback. Treatment compliance increased by 25 per cent in the group receiving explanations. The relationship between explanation quality and compliance was quantified using a simple linear model

$C = \beta_0 + \beta_1 E + \varepsilon$

where $C$ represents the proportion of prescribed actions followed, $E$ is the explanation quality score on a five-point scale, and $\varepsilon$ is the error term. The estimated coefficient $\hat{\beta}_1 = 0.05$ indicates that each one-point increase in perceived explanation quality corresponded to a further five-percentage-point gain in adherence. The model achieved an $R^2$ of 0.41, reflecting a moderate but clinically meaningful association.

**Discussion**

Across diverse healthcare contexts, explainable artificial intelligence has demonstrated clear benefits. Clinician trust improved by up to 30 per cent when AI predictions were accompanied by interpretable explanations. Model accuracy remained close to baseline deep-learning performance, with only slight reductions and manageable computational costs. Patient outcomes also improved, as illustrated by the significant increase in diabetes treatment compliance associated with higher-quality explanations. These findings confirm that integrating explainability into clinical AI systems strengthens both technical performance and the human–AI partnership essential for safe and ethical healthcare delivery.

**CONCLUSION**

Explainable artificial intelligence (XAI) stands at the intersection of technological sophistication and clinical necessity, offering a vital pathway for integrating advanced machine-learning models into routine healthcare. The evidence reviewed in this paper shows that XAI techniques, most notably SHapley Additive exPlanations (SHAP), Local Interpretable Model-Agnostic Explanations (LIME), attention mechanisms, and Gradient-weighted Class Activation Mapping (Grad-CAM), consistently improve transparency without materially eroding predictive power. Across a range of clinical applications, from radiology to electronic health record (EHR) analysis, these methods have been shown to raise clinician trust by as much as 30 per cent and to enhance patient adherence to treatment recommendations by a quarter, all while maintaining near–state-of-the-art accuracy. Such findings confirm that interpretability is not merely a desirable adjunct but a prerequisite for ethical and effective AI deployment in medicine.

Despite these gains, substantial barriers continue to impede widespread adoption. Computational cost remains a prominent concern, particularly for algorithms such as SHAP whose complexity grows exponentially with the number of input features. Real-time clinical environments require explanations that are not only accurate but also generated with low latency, a challenge when dealing with high-dimensional medical data. A further difficulty lies in the absence of universally accepted metrics for explanation quality. Existing evaluations rely heavily on qualitative clinician feedback, which, while valuable, lacks the standardisation needed for regulatory oversight and cross-study comparison. Regulatory expectations themselves are in flux. Frameworks such as the European Union's General Data Protection Regulation mandate a "right to explanation", yet they stop short of defining what constitutes a sufficient explanation, leaving developers uncertain about compliance thresholds and exposing healthcare organisations to legal ambiguity.

Future progress will depend on a multi-pronged strategy. First, hybrid modelling approaches that embed interpretability within the architecture, such as combining transparent components with deep neural networks, offer a promising route to balance accuracy and clarity. Second, the development of robust evaluation frameworks co-designed with clinicians, patients, and regulators will be essential to create metrics that capture both mathematical fidelity and clinical relevance. Third, education and capacity building must not be overlooked: healthcare professionals require training to critically appraise AI outputs and to communicate algorithmic reasoning to patients in plain language. Without such efforts, even the most technically sophisticated explanations may fail to achieve their ultimate goal of informed and trustworthy clinical decision-making.

In sum, XAI provides a critical bridge between the predictive strength of contemporary machine learning and the transparency demanded by medical ethics, patient rights, and regulatory bodies. Sustained collaboration among computer

scientists, clinicians, policymakers, and educators will be necessary to overcome current obstacles and to ensure that explainable AI realises its full potential as a transformative force in healthcare.

**REFERENCES**

Amann, J., Blasimme, A., Vayena, E., Frey, D., & Madai, V. I. (2020). Explainability for artificial intelligence in healthcare: a multidisciplinary perspective. *BMC Medical Informatics and Decision Making*, 20(1), 310.

Bahdanau, D., Cho, K., & Bengio, Y. (2014). Neural machine translation by jointly learning to align and translate. *arXiv preprint* arXiv:1409.0473.

Caruana, R., Lou, Y., Gehrke, J., Koch, P., Sturm, M., & Elhadad, N. (2015). Intelligible models for healthcare: Predicting pneumonia risk and hospital 30-day readmission. *Proceedings of the 21st ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, 1721–1730.

Doshi-Velez, F., & Kim, B. (2017). Towards a rigorous science of interpretable machine learning. *arXiv preprint* arXiv:1702.08608.

Goodman, B., & Flaxman, S. (2017). European Union regulations on algorithmic decision-making and a "right to explanation". *AI Magazine*, 38(3), 50–57.

Holzinger, A., Langs, G., Denk, H., Zatloukal, K., & Müller, H. (2019). Causability and explainability of artificial intelligence in medicine. *Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery*, 9(4), e1312.

Huang, S., Cai, N., Pacheco, P. P., Narrandes, S., Wang, Y., & Xu, W. (2019). Applications of deep learning in precision oncology. *Journal of Hematology & Oncology*, 12(1), 101.

Kim, B., Park, J., & Lee, S. (2023). Explainable AI in radiology: Enhancing clinician trust with visual explanations. *Journal of Medical Imaging*, 10(2), 024501.

LeCun, Y., Bengio, Y., & Hinton, G. (2015). Deep learning. *Nature*, 521(7553), 436–444.

Lee, C., Kim, H., & Park, Y. (2024). Trust in AI-driven EHR diagnostics: The role of explainability. *Health Informatics Journal*, 30(1), 145–156.

Lundberg, S. M., & Lee, S.-I. (2017). A unified approach to interpreting model predictions. *Advances in Neural Information Processing Systems*, 30, 4765–4774.

Molnar, C. (2020). *Interpretable Machine Learning: A Guide for Making Black Box Models Explainable*. Leanpub.

Patel, R., Sharma, A., & Gupta, S. (2024). Impact of explainable AI on patient compliance in diabetes management. *Journal of Clinical Medicine*, 13(3), 789.

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.

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.

Rudin, C. (2019). Stop explaining black box machine learning models for high stakes decisions and use interpretable models instead. *Nature Machine Intelligence*, 1(5), 206–215.

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.

Wiens, J., Saria, S., Sendak, M., Ghassemi, M., Liu, V. X., Doshi-Velez, F., … & Goldenberg, A. (2019). Do no harm: a roadmap for responsible machine learning for health care. *Nature Medicine*, 25(9), 1337–1340.

Zhang, L., Wang, Q., & Chen, X. (2024). Performance evaluation of explainable AI methods in medical imaging. *Medical Image Analysis*, 92, 103045.