Review

Overcoming the Black Box Challenge: Building Trust in Artificial Intelligence Algorithms in Oncology

Technology in Cancer Research & Treatment
Volume 25: 1-21

DOI: 10.1177/15330338261434649
journals.sagepub.com/home/tct

Esther Ugo Alum, PhD[1], Chukwuoyims Kevin Egwu, PhD[2], Vaithiyalingam Subramanian Manjula, PhD[3], Patience Owere Ekpang, PhD[4], Joseph Enyia Ekpang II, PhD[5], Darlington Arinze Echegu, BSc[1,3], Benedict Nnachi Alum, BSc[1,3], and Daniel Ejim Uti, PhD[1]

Abstract

Rising global cancer rates are projected to significantly increase by 2050, highlighting the urgent need for improved scalable prevention, early detection, and personalized therapy tools. Artificial intelligence (AI) has demonstrated significant capabilities in diverse oncology tasks, leveraging high-dimensional data from medical imaging, molecular profiles, and electronic health records for applications in radiology, digital pathology, genomics, prognostication, and treatment selection. Nevertheless, the clinical adoption of most AI systems is still limited by the black box issue, that is, prediction without clear explanation, which, in turn, limits the confidence and accountability of clinicians as well as their ability to communicate with patients. In this review, we searched sources over the years (2015-2025) from PubMed, Scopus, and Web of Science for evidence on explainable AI (XAI) methodologies that may provide greater interpretability and trust in oncologic practice. Local interpretable model-agnostic explanation and Shapley additive explanations (LIME and SHAP) are model-agnostic methods that offer local and global feature attribution and help clinicians to understand the main influential factors behind model predictions. The complementary approaches, such as Gradient-weighted Class Activation Mapping (Grad-CAM), Integrated Gradients and DeepLift, also bring the explainability to image- and genomics-based processes, whereas more recent strategies (eg, Anchors, Prototypical Part Network (ProtoPNet), and contrastive or counterfactual explanations) also focus on enhancing stability and clinical utility. Irrespective of such developments, several issues continue to be experienced, including computational load, inconsistency in explanations, domain transfer, deployment into clinical processes, bias, privacy issues, and changing regulatory requirements. In general, XAI can transform oncology AI to become clinically interpretable, transparent prediction of outcomes, which will make its application safer by adhering to strict validation procedures, human control, and patient-centered communication. By providing a comprehensive and clinically grounded overview, this review aims to support researchers, clinicians, and stakeholders in advancing trustworthy and transparent AI deployment in oncology.

Plain Language Summary Title

Overcoming the Black Box Challenge in Cancer Diagnosis and Care

Plain Language Summary

Cancer cases are expected to rise dramatically in the coming decades, creating an urgent need for better ways to detect and treat the disease. Artificial intelligence (AI) can support doctors by analyzing medical scans, genetic data, and health records to improve early diagnosis and guide treatment decisions. However, many AI tools work like "black boxes"—they provide results without

[1]Department of Research and Publications, Kampala International University, Kampala, Uganda
[2]Department of Business Administration, Faculty of Management Sciences, Alex Ekwueme Federal University, Ndufu-Alike, Nigeria
[3]Department of Computer Science, Kampala International University, Kampala, Uganda
[4]Department of Information Technology, Kampala International University, Kampala, Uganda
[5]Department of Journalism and Media Studies, Kampala International University, Kampala, Uganda

Corresponding Author:
Esther Ugo Alum, Department of Research and Publications, Kampala International University, P. O. Box 20000, Kampala, Uganda.
Emails: esther.alum@kiu.ac.ug; alumesther79@gmail.com

showing how those results were reached. This lack of transparency makes it difficult for doctors and patients to fully trust AI recommendations.

Our review looks at "explainable AI" (XAI), a growing field that focuses on making AI decisions clearer and easier to understand. Methods such as Local Interpretable Model-Agnostic Explanations (LIME) and Shapley Additive Explanations (SHAP) can show which features or risk factors influenced a prediction, while other tools highlight important areas in medical images. These approaches help doctors validate AI findings, communicate more effectively with patients, and build confidence in using AI in real-world cancer care.

We conclude that explainable AI could make cancer treatment safer, more transparent, and more personalized. For this to happen, researchers, clinicians, and policymakers must work together to improve technical methods, set ethical standards, and ensure patients receive clear and understandable explanations.

Keywords
cancer, oncology, artificial intelligence, explainable AI, personalized medicine

Introduction

In the forthcoming decades, cancer incidence is anticipated to rise significantly due to causes such as exposure to pollutants, tobacco use, and unhealthy lifestyles. Recently, Bray et al[1] predicted in a demographics-based study that the global cancer incidence will increase to 35 million cases by 2050, representing a roughly 77% increase from the 2022 level. Nonetheless, Artificial intelligence (AI) could mitigate that increase by enhancing cancer prevention, facilitating early diagnosis, and personalizing cancer treatment.[2] AI tools can analyze extensive datasets, such as medical images, genetic profiles, and electronic health records, enhancing early detection rates and accurately identifying at-risk individuals.[3] Furthermore, AI-driven prediction models facilitate the formulation of treatment plans and drug development, yielding more efficacious medicines and improved patient outcomes.[4] Thus, AI has significantly transformed oncology, enabling earlier cancer detection and personalized treatment. However, AI's clinical integration is hindered by the so-called "black-box" problem, wherein the decision-making process behind AI predictions is opaque and difficult for clinicians to interpret and trust.[5] The lack of transparency in this regard is a major problem for clinicians who are required to comprehend, trust, and discuss AI-produced suggestions with patients.[6] For deep integration of AI into oncology, in which decisions have downstream consequences on patient lives, overcoming the black box challenge is crucial. This narrative review considers the implications of these challenges and outlines avenues to increase transparency and trust in the use of AI systems in oncology. This review explores the potential of XAI methods to enhance AI transparency in oncology, improve clinician trust, and enable more effective patient care.

A number of previous studies have examined XAI in healthcare or oncology contexts, but this literature tends to focus on a specific set of commonly used approaches (mostly SHAP, LIME, and Grad-CAM), mainly focuses on quantitative performance metrics, or lacks sufficient attention to clinical trust and rigorous validation as well as practical constraints of real-world applications. This present review attempts to overcome these shortcomings by providing an integrative synthesis of existing, as well as emerging, XAI approaches and methods: intrinsically interpretable modeling, counterfactual and exemplar-based modeling, global interpretability methods, and pragmatic toolkits, and contextualizing the discussion within the context of workflows and decision-making in the oncology field. For example, Mohamed et al[7] conducted a PRISMA-oriented systematic review of XAI use in diagnosis, prognosis, and therapeutic planning of cancer where the majority of research included the existing formats of explanation and highlighted their shortcomings, including insufficient clinician interaction. Conversely, the current review follows a trust-focused, deployment-based viewpoint, which focuses on human-AI cooperation, is more critical of the dangers to validity, and deals with ethical and regulatory readiness, and considers the less-represented methodological choices and toolchains such as Partial Dependence Plot and Individual Conditional Expectation (PDP-ICE), permutation feature importance, Explain Like I'm 5 (ELI5), and QLattice in addition to the traditional SHAP, LIME, and Grad-CAM strategies. Again, Cui et al[8] focused on the interpretability in radiology and radiation oncology but the present scope includes the full extent of oncology and a multimodal paradigm that goes beyond imaging and includes genomic, radiomic, tabular, and combined analytical pipelines. Although Sadeghi et al[7] present an overarching healthcare synthesis of XAI families, the current review will focus on the issues that are unique to oncology deployment specifically the dataset shift across hospitals and scanners, collinearity in features in radiomics and genomics, and the use of calibration, and offers a prospective framework demonstrating how such challenges can be overcome as the only viable path to trustful oncological AI.

To strengthen transparency despite the narrative orientation, the review is based on a PRISMA -ScR methodology and is supported by a modified checklist that is presented as an appendix. Through the combination of methodological breadth with robust clinical and translational outlook, this narrative review provides a differentiated and prospective framework that will be used to develop transparent, reliable, and clinically significant AI systems in the field of cancer.

Although the most popular techniques applied to tabular, general black-box, and vision-based models are SHAP, LIME, and Grad-CAM, respectively, limiting a review to

these techniques alone risks diminishing important families of explainability techniques. Based on this, the current study takes a larger taxonomy of XAI methods applicable to healthcare decision-making. Besides feature-attribution and saliency approaches, we also have intrinsically interpretable models, example-based explanations, counterfactual and recourse-oriented explanations, concept based and attention based explanations, surrogate and rule-extraction explanations, uncertainty, calibration-based, and causality-informed explanations that affect the way explanations should be interpreted in practice. Incorporating global interpretability, practitioner toolkits, and interpretable-by-design modeling (QLattice) provide a broader and more implementation-oriented reference. This increased coverage is aimed at making the study useful to researchers, clinicians, and other stakeholders and retain the discussion within the framework of practical clinical implementation. The objectives of this review include to:

(i) Thoroughly review the XAI techniques used in oncology, including current and novel techniques.
(ii) Categorize and frame XAI techniques based on methodological underpinnings, and to their applicability to oncology-specific modalities of data.
(iii) Appraise the capabilities, constraints, and clinical utility of the various XAI methods, with special emphasis on reliability, robustness, and interpretability in high-stakes oncologic decision-making.
(iv) Determine the most important research paucities that impede the safe and effective translation of XAI into the regular practice in oncology.
(v) Provide future directions for XAI in oncology, such as collaboration between humans and AI, ethical and regulatory aspects, and both clinical workflow integration.

Through this focused synthesis, the review aims to inform researchers, clinicians, and stakeholders on best practices and unmet needs in building transparent and reliable AI models for oncology.

Methodology

This study was conducted as a narrative review designed to synthesize and contextualize XAI methods in oncology and did not follow a strict systematic review or meta-analysis protocol. The narrative review methodology was used to allow integrating heterogeneous approaches, data forms and application settings conceptually, which would not readily adapt to formal meta-analysis processes or strict systematic review guidelines. This narrative review prioritizes conceptual clarity, methodological breadth, and clinical relevance over exhaustive enumeration, which is appropriate given the heterogeneity and rapid evolution of XAI methods in oncology.

Ensuring Methodological Clarity, Transparency, and Reproducibility

Although formal systematic procedures (eg, protocol registration, duplicate independent screening, quantitative pooling) were not applied, several measures were implemented to ensure transparency and reproducibility for readers. First, the review process was explicitly structured and documented, including information sources, search terms, inclusion rationale, and synthesis strategy. Second, the review mimicked the PRISMA-ScR framework, which guided clear reporting of literature identification, selection rationale, and thematic organization, without implying full compliance with scoping or systematic review standards. Third, all methodological choices and limitations inherent to a narrative review design are explicitly acknowledged. A brief PRISMA-ScR–informed checklist summarizing the reporting elements addressed in this narrative review is provided as Supplemental Material 1 (Supplementary Table S1).

Literature Sources and Search Strategy

A literature search was carried out across reputable scientific databases related to the research field of oncology, biomedical studies, and artificial intelligence including PubMed, Scopus, and Web of Science for peer-reviewed articles published between 2015 and 2025. Search terms included combinations of "artificial intelligence", "explainable AI", "XAI", "oncology", "LIME", "SHAP", "Grad-CAM", "interpretability", and "clinical decision support systems". Priority was given to studies demonstrating clinical relevance, use-case applications in radiology, pathology, genomics, and treatment decision-making. Emerging methods, ethical implications, and patient-centered approaches were also considered. Additional sources were identified through snowballing from references in key papers. Clinical experts including oncologists were also consulted.

Rationale of Study Selection and Inclusion

Only English-language articles that addressed AI interpretability techniques in oncological contexts were included. The inclusion criteria were as follows: (i) the study used machine-learning or deep-learning on oncology-relevant tasks (ie, diagnosis, prognosis, treatment response, risk stratification); and (ii) at least one XAI or interpretability technique was used in the study. They were both the widely used and the emerging XAI techniques in order to safeguard conceptual comprehensiveness. Non-oncology studies that lacked a discrete explainability element and those that were not described methodologically were eliminated. Because of the narrative review format, formal duplicate screening and quantitative eligibility scoring were not done; relevance was determined iteratively depending on domain applicability, methodology contribution and prominence of citations.

Data Extraction and Thematic Synthesis

Qualitative information was extracted on each of the included studies, and includes the type of cancer, the type of data, model type, XAI method(s), level of explanation (local/global), validation method and the strengths and limitations reported. As an alternative to statistical data synthesis, the results were synthesized thematically to compare the XAI approaches to oncology use-cases in order to reveal typical patterns of methodology and also to reveal the gaps in clinical validation and implementation. The findings of this study were presented and discussed simultaneously using sub-sections to enhance flow and clarity.

Scope and Methodological Limitations

Being a narrative review, this study does not claim to cover all the literature available, and does not follow a formal PRISMA-guided selection pipeline. However, by explicitly documenting search sources, inclusion rationale, and synthesis strategy, the review ensures reproducibility of approach while retaining the flexibility necessary to integrate diverse and rapidly evolving XAI methodologies in oncology.

Explainable AI (XAI) Techniques

XAI encompasses various techniques designed to make machine learning (ML) models more interpretable. They clarify what drives a prediction and how model behavior changes across inputs. The drive towards clinically interpretable AI can also be identified in the related fields, including the application of explainable model-driven solutions in the noninvasive diagnosis of clinically relevant portal hypertension,[9] gestational diabetes prediction,[10] and sickle cell diseases.[11]

In the oncology, can serve various purposes: (i) through making predictive outcomes consistent with the known oncologic theory and biological processes; (ii) by making predictive models more auditable and valid; and (iii) by making predictive model output more communicable to the stakeholders.[12] More importantly, explainability can be expressed either locally, in the context of a prediction of a single patient, or on a global scale, across the behavior of the model as a whole. Additionally, explainability can be done post-hoc, by explaining a previously fitted black-box model, or intrinsically, by the use of models which are interpretable in nature.[13] XAI is increasingly used to support interpretability across multimodal pipelines (imaging, radiomics, genomics, and pathology).[7] Here we present these techniques according to their modes of operation.

Local, Model-Agnostic Explanations: LIME and SHAP

LIME offers local interpretability, explaining individual predictions by perturbing inputs and observing the resulting changes in model behavior.[7] LIME is attractive because it can be applied to many model types and provides intuitive feature contribution summaries. However, LIME explanations can be unstable across runs and sensitive to feature correlation and sampling choices; these conditions are common in radiomic and multi-omic oncology data.[14] SHAP, based on cooperative game theory, provides consistent feature attribution, allowing a deeper understanding of how individual features influence a model's output.[15] When evaluating the suitability of XAI techniques for oncology applications, several factors must be considered: effectiveness, computational cost, and clinician accessibility. For instance, SHAP provides robust global and local interpretability, making it suitable for genomic and prognostic models where understanding the contribution of each feature is crucial.[16] However, SHAP is computationally expensive, particularly for models with numerous features.[17] In contrast, LIME, while more efficient computationally, often lacks global interpretability and is more effective in explaining individual predictions than in providing insights into the overall model behavior.[18]

Visual Explanations for Imaging Models: Grad-CAM and Saliency Methods

For image-based oncology tasks, Gradient-weighted Class Activation Mapping (Grad-CAM), often used in deep learning models, generates visual explanations by highlighting image regions critical to the model's decision-making process, proving particularly useful in medical imaging applications such as histopathology, mammography and Computed Tomography (CT) scans.[19] While useful for qualitative inspection, saliency methods like Grad-CAM can be visually persuasive but not necessarily faithful as highlighted regions may shift with preprocessing, architecture changes, or noise, and may reflect artifacts rather than pathology. Therefore, they are best paired with quantitative checks (eg, perturbation/occlusion tests) and external validation across scanners, institutions, and cohorts.[7] In sum, Grad-CAM excels in image-based oncology applications, but its applicability is limited to convolutional neural networks (CNNs), restricting its use in non-image-based data like genomics.[20]

Backpropagation-Based Attribution for Deep Models: Integrated Gradients and DeepLIFT

Integrated Gradients (IG) calculate feature relevance by adding up gradients on a path between a reference input and the target input that is of interest to the model in question. Deep Learning Important FeaTures (DeepLIFT) attribute generates a neuron activation that is used to attribute model output to input feature by comparing it to neuron activation resulting from a given reference baseline.[21] These methods have been effectively applied to a variety of deep learning designs, such as but not restricted to image analyses and they have been shown to produce more consistent attribution maps than gradient based methods, which use raw gradients only. Their interpretability depends greatly on baseline/reference selection and input scaling, which can be non-trivial in clinical settings. Integrated Gradients and

DeepLIFT focus on computing the contributions of each input feature to the model's predictions, offering transparency in tasks such as genomic data analysis or drug response prediction.[22]

Surrogate Models and Rule Extraction (Global Interpretability)

An interpretable model can be trained to provide a high-level view of model behavior and decision boundaries. Global surrogates (eg, training an interpretable model to approximate a black-box model) and rule extraction methods give a high-level view of model behavior and also give a high-level view of decision boundaries. They can be applicable in governance, audit and communication to the stakeholders. Its main shortcoming is that of approximation error: a surrogate might seem interpretable but not faithfully represent the original model, particularly in high-sparsity-density regions or highly complicated interactions.[7] Surrogate models, such as decision trees or linear regression, provide approximations of more complex black-box models, offering clinicians a clearer, albeit simplified, view of AI decision-making.[23] Techniques like Surrogate Models provide a simple, interpretable alternative but may fail to capture the full complexity of deep models, especially in high-dimensional domains like radiomics.[24] Table 1 provides a brief comparison of the various XAI techniques, their advantages, and limitations. Despite their popularity, both LIME and SHAP exhibit limitations that can affect their reliability in clinical oncology. LIME is known to produce unstable explanations, particularly in the presence of highly correlated input features**,** which are common in genomic and radiomic datasets. As a result, the same instance may yield different explanations across repeated runs, reducing clinician confidence.[25] SHAP, while offering both global and local interpretability, is computationally intensive, especially for complex models with many features. Moreover, its attributions can be misleading if the underlying model is poorly calibrated or overfitted, a common issue in healthcare AI.[26] To address these gaps, newer methods such as Anchors, which provide high-precision rule-based explanations; ProtoPNet, which identifies prototypical parts of data that influence classification; and Contrastive Explanation Methods (CEM)**,** which highlight features that differentiate one prediction from another, are being explored. These emerging techniques aim to improve stability, clarity, and clinical relevance, and may play an increasing role in the future of XAI in oncology.

Beyond SHAP, LIME, and Grad-CAM: A Broader Taxonomy of XAI Methods for Healthcare

Within this broader taxonomy, SHAP and LIME remain central for local feature-attribution in tabular and general black-box settings, while Grad-CAM and related saliency approaches remain prominent in deep vision models. Nevertheless, the bigger picture is that no one approach is adequate across modalities and clinical operations; the triangulation of explanations, that is, a combination of feature attributions, counterfactuals and example-based evidence is frequently more justifiable in clinical practice. Although some of these methodologies are yet to be broadly validated or routinely applied to the oncology practice, including them provides a complete landscape of references and provides insight into the future directions as to which the further methodological integration could be made. Importantly, the low adoption is not synonymous to the lack of relevance. In oncology, where clinical decisions are high-stakes, longitudinal, and multimodal, explainability requirements extend beyond model accuracy to include biological plausibility, robustness across cohorts, interpretability across disease stages, and compatibility with clinical workflows. As a result, emerging or underdeveloped XAI tools can take on a central role in addressing the lack of satisfaction in the demands relating to trust calibration, treatment planning, and regulatory approval, despite the early stage of the available evidence.[10,11]

Intrinsically Interpretable Models (Built-in Transparency)

Intrinsically interpretable models can be completely understood without examining the underlying model. Another major substitute to post hoc explainability is the use of models whose design can be interpreted directly. Some examples are the sparse linear or logistic regression, the generalized additive models (including interpretable GAM variants), decision trees, rule lists, and scoring systems. Such models are usually provided with transparent relations of features to outcomes and can eliminate the need to provide retrospective layers of explanation.[34] In clinical practice, the benefits of intrinsically interpretable models can also be seen when clinical accountability and auditability are of primary importance. However, these models can be compromised in terms of predictive performance in very complex tasks and their interpretability can be lost when feature engineering, interactions and high-dimensional inputs are non-trivial.[35]

Case-Based Reasoning and Prototypes (Example-Based Explanation)

The example-based methods explain a prediction with references to similar situations (nearest neighbors), prototype representatives, or most influential training examples. Analogies with previous patients may be easier to understand than abstract weights of features, which, in turn, would help clinicians with face-validity and reviewing cases.[36] Nevertheless, example-based explanations are sensitive to the data quality, their representativeness, and the selected metric of similarity; they can also reveal the privacy threats unintentionally unless they are managed.[37]

Table 1. Comparison of XAI Techniques: Advantages & Limitations.

XAI Technique	Advantages	Limitations	References
Local Interpretable Model-Agnostic Explanations (LIME)	- Model-agnostic, can be applied to any classifier. - Provides local interpretability (explains individual predictions). - Allows for better understanding of complex models.	- Requires sampling, which can be computationally expensive. - Does not always provide global interpretability.	27
Shapley Additive Explanations (SHAP)	- Considers the impact of all features in the prediction. - Provides both global and local interpretability. - Guarantees consistency and fairness in feature importance.	- Computationally expensive, especially for complex models with many features. - Can be challenging to implement on large datasets.	28,29
Gradient-weighted Class Activation Mapping (Grad-CAM)	- Provides visual explanations by highlighting important image regions. - Useful in image-based models like CNNs. - Enables easy interpretation of model decisions in medical imaging.	- Limited to convolutional neural networks (CNNs). - May struggle with models that do not use convolutional layers.	30,31
Integrated Gradients	- Simple and interpretable method. - Quantifies the importance of each feature on the final prediction. - Applicable to various neural network architectures.	- Requires selecting a baseline for comparison, which can be challenging. - Can be computationally intensive for large models.	32
Deep Learning Important FeaTures (DeepLIFT)	- Efficient method for explaining deep neural networks. - Provides attributions for input features in comparison to a reference. - Faster and less computationally expensive than other techniques like SHAP.	- Limited to deep learning models. - May not always align with human intuition for feature importance.	32,33
Surrogate Models (eg, Decision Trees, Linear Models)	- Simple and interpretable models, easy to visualize. - Can be used to approximate the behavior of complex black-box models.	- May not capture the complexity of the original model's decision-making. - Performance of the surrogate model can vary depending on the data.	27

Counterfactual Explanations and Algorithmic Recourse (Actionability)

Counterfactual explanations explain what would happen to a model with minimal perturbations to the input. Counterfactuals in clinical decision support can be used to increase actionability, making predictions connected to the possible intervention or monitoring plan.[38] However, counterfactual validity can be jeopardized when the proposed changes are clinically impossible to implement (eg, fixed features), they do not respect causal restrictions, or they are not consistent with established care pathways.[39] In line with that, clinically based constraints and consequential logic are necessitated when giving counterfactuals in the health context.

Concept-Based Explanation and High-Level Interpretability

Concept based approaches seek to align model reasoning to match human interpretable clinical concepts (such as edema patterns, some radiographic appearance, laboratory syndromes) as opposed to raw pixels or individual features. Such correspondence can fill the semantic disparity between model internals and clinician reasoning.[40] Concept definition and the need to ensure that the model really makes use of the concepts and not just spurious correlates is a continuous challenge, especially when dealing with different populations and imaging equipment.

Attention Mechanisms: Useful (Non-Explanatory) Mechanisms

Attention weights and attention maps are commonly used as natural-language and imaging tasks often used as an intuitive attribute of attention. Regions or tokens related to model processing can be emphasized by attention and qualitative review could be aided by it. Attention, however, does not imply fidelity to the importance of causality, and its interpretability is determined by the architecture used and the validation methods.[41] Attention based visualizations must therefore be introduced with care and where possible with faithfulness checks or additional techniques (eg, perturbation tests, counterfactuals).

Uncertainty, Calibration, and Causality: Interpretability in Context

Explainability is most relevant when it is combined with a trusted uncertainty quantification and calibration since clinicians should be able to know not only the reasoning behind a prediction but also the confidence in a prediction and what conditions may cause a prediction to fail.[42] Furthermore, several of the outputs of explanation are meant to be associational, not causal, lack of causal support causes explanations to strengthen confounding or data artifacts. Thus, XAI results are to be discussed together with calibration measures, external validation, sub group results and clinical plausibility checks.[43]

The Permutation Feature Importance (PFI)

PFI is used to determine feature relevance through the quantification of the predictive performance loss after feature value permutation. Due to its model-agnostic character and computational simplicity, PFI is commonly used in fast global interpretability and debugging of a model. However, PFI can give biased estimates when predictors are correlated and can miss the significance of variables that share information with others.[44] Therefore, careful interpretation and sensitivity analysis are required, particularly when working with clinical data that often has collinearity.

Partial Dependence Plot (PDP) and Individual Conditional Expectation (ICE)

PDPs represent the average connection between a predictor and model result whereas ICE graphs illustrate instance-specific response curves. The PDP–ICE approach is especially beneficial in oncological settings with either risk scores, dose response curves, or continuous biomarkers because it is capable of revealing non-linear effects and interactions that can be obscure in purely local attribution approaches.[45] However, PDPs are misleading in correlated feature space or extrapolation in sparsely populated feature space, ICE plots can be useful in detecting heterogeneity but can still be used to show association as opposed to causality.

Explain Like I'm 5 (ELI5)

ELI5 is an interpretability pragmatic toolkit, which offers various explanatory results, such as model weight inspection of linear models, feature contribution elucidations of chosen models, and feature importance through permutation. Its major strength is its usability, which allows applied researchers to produce transparent summaries during the process of model development and reporting.[46] ELI5 should, however, be seen as an interface tool, not as a complete XAI methodology, the quality of its explanations depends upon the underlying estimator, and its assumptions.

QLattice

QLattice is a symbolic modeling approach wherein the target of the modeling is small functional representations, such as concise equations or graphs, predicting the relationships thus giving intrinsically interpretable models. The symbolic models may be used to supplement the opaque predictors in situations where the decision logic required by stakeholders is straightforward and auditable (like in clinical risk stratification), and provides transparent alternatives or high-level hypotheses about the association between biomarkers.[47] Just like any other model-search method, QLattice is prone to overfitting, cohort-specific instability, and poor performance in complex tasks, hence external validation and clinical plausibility tests are inseparable.

Historical Development of XAI: A Timeline Perspective

The evolution of XAI has been instrumental in addressing the interpretability challenge posed by black-box AI models. These milestones highlight how the field has progressed from foundational techniques to domain-specific adaptations that enhance clinical trust and usability, as outlined below.

(i) *2016 – Introduction of LIME*: Ribeiro et al[27] introduced LIME, a model-agnostic technique that explains individual predictions by locally approximating the black-box model. It marked the first widely adopted effort to offer interpretability across model types in healthcare.

(ii) *2017 – Emergence of SHAP*: Lundberg and Lee[28] introduced SHAP, which uses game theory to fairly distribute feature importance. Its ability to offer both global and local explanations made it a favored tool in oncology applications, such as genomic mutation analysis.

(iii) *2017 – Grad-CAM Developed for Visual Explanations*: Grad-CAM became a key method for producing visual heatmaps in CNNs, revolutionizing explainability in medical imaging—particularly in radiology and pathology.[48]

(iv) *2019 – Integration of XAI in Cancer Classification Models*: Researchers began integrating LIME and SHAP into neural networks for breast cancer classification and tumor grading from histopathology images, providing visual justifications for model decisions.[49]

(v) *2021–2023 – Deep Integration of XAI in Multi-modal Oncology Models*: With the surge in multi-modal data (genomic, imaging, and clinical records), hybrid and ensemble methods leveraging SHAP and Grad-CAM started appearing in radiogenomics and treatment response prediction studies.[50]

(vi) *2024 – Advances in Real-Time XAI and Federated Learning*: New developments, such as instance-specific SHAP (InstanceSHAP) and federated XAI

strategies, are emerging to meet the growing demand for interpretable models trained across decentralized oncology datasets.[13]

The Black Box Phenomenon

Black boxes are AI systems whose internal decision logic is difficult to understand. This lack of explainability in oncology can raise questions about AI recommendations, particularly when a life-altering treatment choice is at stake.[7] This lack of transparency in the oncological field is not an unimportant inconvenience; instead, it has significant clinical confidence, patient safety, responsibility, and regulatory acceptability implications. Even a model that has a high predictive accuracy can fail to explain the reason behind its findings (such as the reason to categorize a lesion as a malignancy, place a patient in a high-risk category, or to predict a response to treatment). Without clear thinking clinicians might not be in a position to determine whether a prediction is biologically plausible, is because of spurious association, or it is due to latent confounding.[51]

There are various elements that cause the black-box phenomenon in oncology AI. The first is the complexity of models: deep neural networks are trained on high-dimensional nonlinear representations, which are hard to represent as rules or relationships understandable by humans.[52] Second, heterogeneity and high dimensionality of the data: oncology processes actively combine imaging (CT, magnetic resonance imaging MRI, histopathology) variables, radiomics, genomics and electronic health record (EHR) variables, thus forming feature space in which correlated predictors and latent structure make interpretation difficult.[53] Third, dataset artifacts, shortcut learning: models can learn based on non-clinical signals (scanner specific patterns, staining variability, site specific practices, or documentation habits) that are correlated with outcomes but not biological aspects of the disease.[54] Fourth, distribution shift: performance and explanations may differ between institutions due to scanner differences, protocols, patient population or change in guidelines and the behavior would be unpredictable at deployment.[55]

The clinical risks of oncology AI black-box are well known. The absence of interpretability may lead to automation bias where clinicians trust model outputs more to an extent of uncertainty or being incorrect, or may use potentially useful tools less, due to distrust. It may also hinder the analysis of errors and quality assurance, making it hard to diagnose failure modes (such as poor performance in particular subgroups), and it makes medico-legal accountability hard, since the decisions that affect diagnosis and treatment need justifiable reasons. This means that overcoming the black box requires not only correct models, but also information that the models are correct, stable and interpretable clinically in a manner that facilitates safe decision-making.

Evidence abound suggesting that without a clear understanding of how AI reaches its conclusions, healthcare professions may be reluctant to trust tools driven by these technologies, conversely undermining their clinical utility.[5] Oncology in particular tends to have a high stake for such decisions because they are inherently complex and require multidisciplinary teams to evaluate large volumes of patient data. The inability to interpret AI output can limit collaborative decision-making by prompting team members to question the AI recommendations.[56] Patients may also exhibit skepticism towards technology they are unfamiliar with. The growing degree of complexity of medical data has rendered AI a compelling method for analysis and decision-making in oncology.[57] The "Black Box" dilemma, characterized by restricted interpretability and explanatory capacity, impedes the acceptability of AI in clinical practice.[57] In response to this issue, researchers have established XAI frameworks like LIME and SHAP to clarify complex models and facilitate their integration into medical procedures.[58] Figure 1 is a simple illustration of black box and XAI processes. Many oncology applications, including breast cancer classification[59] and Alzheimer's disease detection,[60] have made use of LIME and SHAP. These approaches provide advantages, including the assessment of feature significance, the visualization of correlations, and the evaluation of individual predictions.[58] Some researchers opine that opaque decisions violate physicians' ethical obligations, while others maintain that such decisions are prevalent in medicine and that empirical validation of accuracy may outweigh the necessity for explanations.[61] Notwithstanding obstacles in research, regulation, and privacy, black-box medicine presents significant advantages in personalized healthcare by utilizing advanced algorithms to analyze extensive health statistics.[62]

The Need for Explainable AI (XAI) in Oncology

AI algorithms often produce highly accurate but nonintuitive outputs, particularly in deep learning models due to their complex nature. However, the decision-making pathways in these models are not readily accessible, and despite their ability to analyze millions of data points, including medical images and genomic data, they are limited to providing answers to specific questions.[63] This opacity presents a challenge in oncology, as it necessitates the creation of individual treatment plans that require clear rationales for each decision. To tackle this issue, a promising approach is XAI, which intends to make AI models more interpretable and transparent. This helps oncologists comprehend the rationale behind a model's conclusions, builds trust, and lays the foundation for adopting AI-driven insights.[64]

The application of XAI in oncology spans several domains, improving both the interpretability and the clinical utility of AI-driven decision-making systems. Table 2 summarizes the applications of XAI across diverse oncology domains. In medical imaging, techniques like Grad-CAM are increasingly used to explain AI-based detection of breast cancer from mammograms or lung cancer from CT scans. By highlighting areas of the image deemed critical by the model, clinicians can validate AI's findings and make more informed decisions.[19,65] In pathology, LIME has been employed to explain AI's tumor

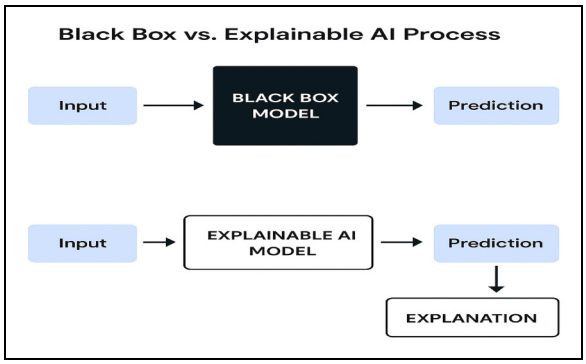

Figure 1. A simple illustration of black box and XAI processes.

classification decisions from histopathology slides, ensuring that pathologists can trust AI suggestions when selecting treatment options.[66] In genomics, SHAP has helped interpret AI models used in identifying cancer-associated mutations, allowing clinicians to understand the rationale behind AI's genetic predictions.[67] Finally, radiomics, which involves extracting large amounts of quantitative features from medical images, has benefited from integrated gradients to clarify how these features contribute to treatment response predictions in cancer patients.[68] Recent studies underscore the increasing significance of XAI in oncology and wider medical fields.[7] Multiple oncology areas, including prognostication, diagnosis, and treatment selection, employ XAI frameworks like LIME and SHAP to enhance interpretability.[58] These model-agnostic methodologies assist in translating complicated machine learning models into comprehensible visual representations, hence promoting their integration into clinical procedures. Researchers have used LIME and SHAP to clarify neural network predictions in breast cancer classification, providing valuable insights into characteristic significance and input-output correlations.[59] Comparable applications have been noted in the identification of Alzheimer's disease, wherein XAI frameworks augment the accuracy of clinical decision support systems.[60] The above tools give quantitative visualizations and comprehensible rules, enhancing the understandability of complex models for healthcare practitioners and potentially augmenting patient outcomes. Medical decision support systems primarily utilize XAI approaches, with deep learning models being the most prevalent.[69] In oncology, XAI elucidates complex machine learning models, enhancing trust among doctors for personalized treatment approaches.[70] XAI is applicable in various medical areas, including radiology, pathology, and cardiology, improving transparency in AI-generated diagnosis and treatment suggestions.[71] In cardiology, XAI techniques are applied in heart disease prediction, particularly in diagnosing arrhythmias or coronary artery disease. Comparative analyses indicate that LIME attains superior performance in separability metrics, whereas SHAP exhibits the most rapid explanation generation.[72] The implementation of XAI in healthcare might enhance doctors' comprehension and confidence in complicated models, hence potentially promoting their integration into medical procedures.[58,60] Despite advancements, obstacles persist in establishing user trust and ethical data utilization.

Strategies for Enhancing AI Transparency and Explainability

There is much promise in using artificial intelligence (AI) to improve oncology to enhance cancer diagnosis, prognosis, and treatment personalization. However, we often criticize AI models, including deep learning approaches, for being "black boxes," meaning we are unable to understand how these models make decisions. Again, transparency and explainability issues associated with AI recommendations pose challenges for clinicians and patients to trust them. To encourage the adoption of AI in oncology, it is crucial to go beyond merely improving predictive accuracy and instead focus on developing algorithms that are easy to understand. To reach this goal, we can employ the following strategies:

(i) *Integrating Feature Importance Analysis*: Feature importance ranking, saliency maps, and attention mechanisms, among other techniques, can also be used to identify which inputs (eg, imaging features or genetic markers) matter most in an AI model's decision. Identifying the features or data points that significantly influence an AI model's decision can enhance its interpretability.[7] For example, in the fields of radiomics, saliency maps can identify salient regions whose presence in the medical image helped the AI make the original diagnosis.[73] There are also novel architectures that enable researchers to make not only predictions but yield explanations for their decisions as well. A prime example of this is a recent study by Hassan et al[74] that showed how a deep learning model can simultaneously identify retinal disease biomarkers and present a "map" to explain its diagnostic reasoning. This transparency can potentially empower clinicians to validate their findings against their clinical judgment.

(ii) *Employing Model-Agnostic (Post-hoc Explanation) Methods*: Post-hoc approaches employ AI models to produce explanations that align with the decisions made by the AI model. Some techniques, such as LIME and SHAP, can explain AI systems after they have been built without changing the structure of the models themselves. These methods make it easier to understand complicated models, which makes predictions easier for oncologists to understand. Eventually, they understand why certain treatment suggestions or risk assessments are made.[75]

(iii) *Developing Hybrid Models*: Machine learning models can combine with traditional statistical methods to form accurate and explainable systems. Hybrid models combine the strengths of deep learning in pattern recognition with the transparency of

Table 2. Applications of XAI Across Oncology Domains.

Oncology Domain	XAI Technique (s) Applied	Applications	Key Benefits	Challenges	References
Medical Imaging (Radiology)	- LIME - Grad-CAM - SHAP	- Breast cancer detection via mammograms. - Lung cancer detection via CT scans.	- Improves interpretability of image-based models. - Enhances clinician trust by identifying critical image features.	- Limited by image quality and resolution. - Struggles with complex multi-class classification	27,30,31
Pathology	- LIME - SHAP	- Tumor classification from histopathology slides. - Identifying metastatic cancer.	- Enables pathologists to understand model decisions. - Provides feature importance, such as cell morphology.	- High dimensionality in slide images may affect performance. - Model interpretability is still a challenge for large datasets.	27,29,33
Genomics (Precision Oncology)	- SHAP - Integrated Gradients	- Identifying mutations linked to cancer. - Biomarker discovery for personalized therapies.	- Increases transparency in genomic predictions. - Helps in selecting treatment based on genetic profiles.	- Difficulty in handling large genomic datasets. - Potential for biased models due to incomplete training data.	29,31,32
Prognostication	- LIME - Surrogate Models	- Predicting patient survival outcomes. - Forecasting recurrence risk after treatment.	- Facilitates personalized prognostic assessments. - Allows clinicians to visualize factors influencing predictions.	- Models may not generalize across different patient populations. - Risk of oversimplifying complex prognostic factors.	27,33
Treatment Selection	- SHAP - Grad-CAM	- Selecting appropriate chemotherapy or immunotherapy regimens. - Personalized radiation therapy planning.	- Provides clear rationale for treatment choices. - Enhances clinician-patient communication.	- Difficulty in creating interpretable models for highly individualized treatments.	29,30
Drug Discovery & Repurposing	- LIME - SHAP - DeepLIFT	- Identifying potential drug candidates. - Repurposing existing drugs for cancer treatment.	- Accelerates the drug discovery process. - Reduces the time to identify effective treatments.	- Lack of transparency in model training data. - Difficulty in translating model predictions to clinical trials.	27,32
Clinical Trials	- LIME - SHAP	- Identifying suitable candidates for clinical trials. - Predicting patient response to trial interventions.	- Improves the efficiency of clinical trial recruitment. - Enables better patient stratification.	- Ethical concerns with using AI in critical decision-making. - Risk of bias in trial eligibility predictions.	27,32
Radiomics	- SHAP - Integrated Gradients	- Analyzing tumor heterogeneity. - Predicting treatment responses based on radiomic features.	- Helps in early cancer detection and treatment response prediction. - Improves the precision of radiomic analysis.	- High computational cost. - Complexity in integrating radiomic features with clinical data.	29,32,33

conventional approaches like logistic regression.[63] Sometimes, choosing inherently interpretable models, like decision trees or logistic regression, over complex deep neural networks is more beneficial. Even though these simpler models may not always achieve the highest accuracy, they serve as more straightforward explanations of decisions, thereby enhancing transparency in clinical applications.[7]

(iv) *Integrating Human Expertise*: Including clinicians in the AI decision-making process ensures the pairing of AI outputs with expert knowledge. Using AI predictions gives clinicians a glimpse into why the AI is making the recommendations it does, as well as assessing whether the recommendations would actually agree with clinical intuition.[76] Researchers from Stanford and the University of Washington have developed an auditing framework that uses human expertise and generative AI to evaluate classifier algorithms. This synergy can be useful to spot for biases or spurious correlations in AI output.[77]

(v) *Model Validation and Regulatory Standards*: First and foremost, we must delineate clear expectations

and procedures for the development and deployment of AI in healthcare settings. Rigorous validation of AI models on large, diverse datasets is thus required to ensure their reliability across different patient populations.[78] Furthermore, by adhering to emerging regulatory frameworks and standards for AI in healthcare, we can develop explainable algorithms for the oncology field. This would aid in fostering trust between staff and patients, while also upholding the importance of patient safety.[7]

(vi) *Fostering a Culture of Collaboration*: Interdisciplinary collaboration among oncologists, ethicists, and data scientists is crucial for an inclusive understanding of AI systems. Clinicians may gain more confidence in employing these technologies through consistent workshops and training sessions focused on AI tools.[79]

(vii) *Building Trust Through Validation and Clinical Trials*: Although transparency is crucial for fostering confidence, it is insufficient on its own to guarantee the dependability of AI algorithms in practical cancer contexts; comprehensive validation and clinical trials are essential to confirm the validity of these algorithms.[80] External validation of AI models using robust and diverse datasets helps confirm the generalizability of these models.[81] Furthermore, employing AI as a decision-support mechanism in clinical trials rather than as a substitute for human discernment can enhance oncologists' trust. This enables doctors to participate in the creation of AI models, ensuring that the emerging AI addresses the practical needs and concerns of the oncology community.[82]

(viii) *Addressing Ethical Challenges*: The ethics of artificial intelligence in cancer must not be disregarded. Artificial intelligence algorithms cause significant concern because training datasets may not adequately represent all patient attributes. To eliminate inequities in cancer care, it is essential to overcome these biases by utilizing diverse and representative datasets.[83] Regulatory organizations must develop clear criteria for AI use in healthcare settings by prioritizing transparency, patient safety, and full accountability.[84]

Enhancing Patient Trust and Communication in XAI-Assisted Oncology

While most XAI research focuses on clinician interpretability, patient trust in AI-assisted care is equally critical. In oncology, where treatment decisions carry profound emotional and physical consequences, patients must understand and feel confident in the tools guiding their care. However, current XAI outputs, such as SHAP plots or Grad-CAM heatmaps, are often too technical for non-expert audiences. As AI-supported decisions increasingly influence diagnosis and treatment options, patients must be able to comprehend the rationale behind these recommendations to make informed choices and maintain confidence in their care.[85]

(i) *Patient-Friendly Explanations*: Traditional XAI outputs, such as SHAP plots or Grad-CAM heatmaps, are often too technical for patients to interpret. To bridge this gap, AI systems must generate simplified, layperson-level explanations tailored to patient literacy.[86] For example, rather than presenting a heatmap, an AI system could summarize its output as: "The model recommends treatment A because your imaging results and tumor markers resemble those of patients who responded well to this therapy."

(ii) *Clinicians Communication Strategies*: Clinicians serve as intermediaries between patients and AI-driven systems. Therefore, effective communication training is essential. Clinicians should be trained to translate XAI outputs into understandable terms, using analogies (eg, "like a second opinion that shows its reasoning") and visual aids.[87] Shared decision-making tools could incorporate interactive XAI visualizations with "what-if" scenarios to demonstrate how changes in patient features might affect predictions.[88] Tailoring the level of explanation based on the patient's health literacy and preferences is very crucial.[89]

(iii) *Building Trust Through Transparency*: Studies show that patients are more likely to trust AI decisions when they are accompanied by transparent rationales and endorsed by their healthcare provider. Thus, clinicians play a key role in mediating the AI-human trust relationship**,** and XAI must be designed to support, not bypass, this human interaction. Incorporating patient-facing XAI interfaces and prioritizing explainability from both provider and patient perspectives will be essential for ethical, equitable, and widespread adoption of AI in oncology.[90] Emerging research suggests patients are generally open to AI assistance in their care but demand transparency and clinician oversight. A study by Longoni et al[91] found that patients are more likely to accept AI-guided decisions when they are paired with clear, human-interpreted explanations. These findings highlight the need for human-centered XAI interfaces that support trust-building through transparency, empathy, and shared understanding. Ultimately, building patient trust in AI-assisted oncology will require not just accurate models, but interpretable and relatable explanations communicated through trusted clinical relationships.

Critical Appraisal of XAI in Oncology: from Plausible Explanations to Clinically Trustworthy Evidence

Even though XAI is frequently introduced as a solution to the problem of the black box, having an explanation are not

essentially tantamount to truth, causality, or clinical usefulness. In oncology, in which decision-making involves high stakes and extends over time, that is, screening, diagnosis, choice of therapy, and response monitoring, an explanation must satisfy at least three conditions to be associated with clinical relevance: (i) faithfulness, which means that the interpretation of the underlying decision logic of the model is accurately represented; (ii) stability, which means that the interpretation remains consistent under minor perturbations; and (iii) clinical utility, which means congruence with oncologic reasoning and support of safe clinical actions.[92] Many published oncology researches give graphic or visually stimulating descriptions, eg, heatmaps or ranked-feature visualizations, but do not formally certify these properties, and thus run the risk of deploying a model that is interpretable, but also insecure or misleading.[12,93]

Method-Specific Limitations and Failure Modes in Oncology

Two of the most popular model-agnostic feature attribution methods, SHAP and LIME, susceptible to dependance structures and are unstable.[94] Oncology datasets often have correlated predictors, eg, radiomic texture families, multi-omic signatures, and correlated laboratory panels. In these correlated environments, feature attribution might be non-unique, that is, several correlated variables can share predictive credit and the resulting importance scores can change significantly across resampling plans, cross-validation folds, or pre-processing choices.[95] LIME additionally depends on local perturbation schemes, kernel bandwidth and surrogate model choice; these design options may produce different explanations of the same patient, thus undermining clinician trust.[14] SHAP may give local and global attributions, but its empirical accuracy is dependent upon the calibration of the model, preprocessing of the data and the feature dependance, which are often not reported in clinical AI studies.[96]

Grad-CAM and saliency mapping methods tend to generate highly appealing visualizations, but their empirical validity is low. In the field of oncologic imaging (such as histopathology, CT/MRI, and mammography), saliency maps are often viewed as reflecting the spatial areas that were used in model decision-making. Such emphasized areas, however, can be due to non-representative artifacts, like scanner markers, staining artifacts, or compression patterns, and thus reflect shortcut learning and not true pathological signals.[97,98] Furthermore, saliency representations are prone to the difference caused by different model architectures or preprocessing pipelines. Without quantitative measures of faithfulness, (such as deletion and insertion curves, occlusions sensitivity, and perturbation based measurements), saliency maps are to be considered as hypothesis-generating devices but not as conclusive proof of biologically based interpretations.

Global approaches like PFI and PDP-ICE can provide inaccurate results under correlation and extrapolation of features. Practically, PFI can underestimate the relative role of features which share information and overestimate role in case there is information leakage or proxy variables.[99] PDP assumes covariates to be marginally independent and may give spurious effects where not empirically supported; ICE, in contrast to being useful in the revelation of heterogeneity, is purely associational.[100] These methodological limitations are presumed to be especially relevant to the oncology field, where clinical decision-making can be directly influenced by interpretation of the importance of biomarkers. Based on this, diagnostic checks should be used to supplement effect plots with sufficient data support (data density), sensitivity to collinearity, and consistency across clinical meaningful subgroups.

The inherent characteristics of surrogates and rule extraction are the exchange of interpretability with approximation error. Surrogate models can provide useful global summaries that can be used to govern and communicate with stakeholders, but create a false impression of understanding in cases where surrogate fidelity is not measured.[101] In the oncology, surrogates should be evaluated on fidelity at local clinical areas (like borderline-risk situations, treatment-threshold situations) rather than using aggregate measures of agreement.

Why interpretability alone does not guarantee trust: calibration, uncertainty, and dataset shift

Clinical trust is undermined less by the absence of explanations than by unreliable performance under real-world conditions. Datasheet shift including scanner variation, staining protocols, site-specific case mix and changing guidelines are regular occurrences in oncology predictive models, and can change both the predictions and the explanations attached to them. Therefore, explanations should be viewed together with estimates of calibration and uncertainty. A model with a good explanation but poor calibration can do more harm than one with a poor explanation but good calibration. The most justifiable deployment strategy in oncology is consequently one that simultaneously provides explanation, confidence and external validation (such as the factors that motivated the prediction, the confidence of the model, and the situations where the model is known to fail).[102]

Clinical Utility and Human Factors: Avoiding Automation Bias

Even explanations of high quality may be harmful when they lead to automation bias, that is, overreliance on the output of the model when clinicians defer conclusion. Where necessary, explanations in oncology tumor boards and workflow-based environments must be crafted to promote the calibrated trust and not persuasiveness. This requires a more human-based evaluative method: exploring the ways in which clinicians in different positions make sense of explanations, evaluating the effect of explanations on diagnostic accuracy and time-to-decision, and establishing whether explanations reduce

or increase inequalities in patient subgroups.[103] The explanations must also be practical: where possible, they should also connect the model outputs with alternatives that are clinically viable (such as further imaging or confirmatory testing) but avoid limitations such as immutable properties and causal plausibility.

Threats to Validity in Oncology XAI Studies

Threats to internal validity: Leakage of outcome-proximal variables (like treatment codes), confounding due to site-specific clinical practices, and failure to report preprocessing procedures, can produce inflated performance and misleading explanations.

Threats to construct validity: Explanatory outputs are often interpreted as a ground truth, despite the inherently associational quality of their outcomes; without an interpretative framework that is conscious of the concept of causality, the results of such explanations may serve to strengthen spurious correlations or proxy variables.

Threats of external validity: Lack of external, temporal validation together with ineffective assessment of subgroup generalizability and domain drift between different scanners or institutional contexts can interfere with predictive performance and the accuracy of the related explanations.

Threats of interpretability: Explanation uncertainty, with the addition of perturbations, such as resampling variation, additive noise, feature correlation, and lack of faithfulness testing can yield explanations that are plausible but not faithful to model reasoning.

Threats of human factors: Unsafe deployment due to automation bias, false interpretation of heatmaps, and lack of clinician training can lead to unsafe use even when technical explanations are available.

Ethical, Legal, and Social Implications of XAI in Oncology

While XAI improves transparency in clinical decision-making, its application in oncology raises significant ethical, legal, and social challenges that must be addressed to ensure responsible deployment and equitable patient outcomes.

(i) *Bias and fairness*: AI models, particularly those trained on non-representative datasets, may inadvertently perpetuate healthcare disparities by providing suboptimal predictions for underrepresented patient populations. For instance, skin lesion classifiers trained primarily on images of light-skinned individuals have shown decreased accuracy in detecting melanoma in darker skin tones, leading to diagnostic bias.[104] In oncology, similar risks arise if tumor histology models are developed on data from a single ethnicity or gender.[105] To mitigate this, XAI techniques can identify biases by visualizing how specific patient attributes influence predictions, ensuring equitable healthcare delivery.[106] However, unless datasets are balanced and diverse, even interpretable models can reinforce systemic biases.

(ii) *Data privacy and Security*: AI in oncology typically involves processing highly sensitive data, including imaging, genomic, and electronic health records. To safeguard patient confidentiality, XAI approaches must integrate robust data anonymization techniques, ensuring that personal data is protected without sacrificing model performance.[107] Techniques such as differential privacy**,** homomorphic encryption**,** and federated learning can allow model training without directly accessing patient-level data.[108] For example, federated learning frameworks enable decentralized training across multiple hospitals, keeping data local while sharing model updates. When combined with interpretable models, such architectures protect privacy while still providing transparent insights into model behavior.[109]

(iii) *Regulatory and governance frameworks*: The deployment of XAI in clinical oncology must adhere to emerging regulatory and governance frameworks**.** Regulatory bodies must develop guidelines to assess the validity and reliability of AI systems, ensuring they meet safety and ethical standards before widespread implementation in clinical practice.[84] Current regulatory pathways for AI such as the FDA's Software as a Medical Device (SaMD) guidelines and the EU's Artificial Intelligence Act emphasize safety, accountability, and human oversight.[110] However, these frameworks are still evolving and often lack specific provisions for evaluating the explainability of AI models. In oncology, where clinical decisions may involve life-altering treatments, XAI systems must meet higher standards for reliability and interpretability. Regulatory bodies may require new evaluation criteria focused on explanation quality, clinical relevance, and user comprehension.[86] Including explainability audits as part of AI approval and post-market surveillance could ensure compliance with ethical standards.

(iv) *Human Expertise hurdles*: Clinicians face information overload, and complex visual outputs from methods like SHAP or Grad-CAM can be difficult to interpret within time-limited settings.[111] Most clinical systems lack seamless integration with XAI tools, and oncology professionals often lack formal training to use them effectively. Interpretability needs also vary by specialty, complicating tool design. Moreover, AI explanations must align with clinical priorities to be useful.[112] Addressing these hurdles requires clinician-centered design, adaptable explanation formats, and integration with existing health IT systems to ensure practical adoption and trust in XAI.

Challenges in Designing Clinical Trials for XAI in Oncology

While validation and clinical trials are critical to the clinical integration of AI in oncology, XAI-specific systems introduce novel and complex challenges that differ from traditional AI trials. Designing robust trials for explainable models requires not only assessing predictive performance but also evaluating interpretability, clinician acceptance, and patient safety in real-world workflows. Addressing these challenges will require interdisciplinary coordination between trialists, oncologists, AI developers, human factors experts, and bioethicists. The creation of consensus guidelines for XAI evaluation in oncology which are analogous to CONSORT-AI or SPIRIT-AI for general AI is a critical next step toward trustworthy clinical validation.

(i) *Defining Evaluation Metrics for Interpretability*: Standard clinical trial outcomes such as sensitivity, specificity, or progression-free survival do not adequately capture the value of model explanations. Trials involving XAI systems must include new metrics, such as explanation fidelity (how well explanations align with model logic), usability scores by clinicians, and trust indices. However, these measures remain non-standardized, limiting their regulatory acceptance.[113]

(ii) *Complexity of Multi-Stakeholder Endpoints*: XAI systems are expected to satisfy multiple stakeholders: clinicians (interpretability), patients (understandability), and regulators (safety and accountability). Designing trials that measure outcomes for all groups adds complexity, requiring qualitative surveys, cognitive task analysis, and patient satisfaction assessments alongside traditional clinical endpoints.[114]

(iii) *Risk of Cognitive Overload and Misuse*: In trials, exposing clinicians to detailed explanations may backfire, leading to over-reliance on AI, misinterpretation of heatmaps, or distraction from critical clinical reasoning.[115] Determining the appropriate level of detail and the best interface for presenting XAI outputs must be rigorously tested in simulated and live settings.

(iv) *Trial Design Heterogeneity*: There is a lack of consensus on how to structure XAI trials. Should they be randomized controlled trials (RCTs), observational evaluations, or implementation science studies[116]? Hybrid designs may be required, such as cluster-RCTs with human-AI interaction arms or crossover trials comparing XAI versus black-box AI systems with human feedback loops.

(v) *Regulatory Ambiguity and Ethics*: Existing regulatory frameworks (eg, FDA SaMD or EU AI Act) do not yet explicitly address how to assess the interpretability of AI models. Moreover, ethical questions arise in trials where AI influences high-stakes decisions like cancer prognosis, especially if XAI explanations are later found misleading.[117] Pre-registration protocols must clearly define how explanations will be used, disclosed, and evaluated.

(vi) *Integration with Electronic Health Records (EHRs)*: To mimic real-world clinical environments, XAI systems in trials must be embedded within oncology information systems and EHRs. However, interoperability issues, delays in model output rendering, and lack of clinician training in XAI interfaces can compromise trial integrity and data reliability.[86]

Research Gaps in XAI for Oncology

Despite the fast development of XAI, its introduction in the existing literature has a number of significant gaps that slow down its reliable introduction into the daily routine of oncological practice. These gaps highlight that progress in oncology-focused XAI must extend beyond algorithmic transparency toward clinically validated, human-centered, and ethically governed systems. To begin with, there is a gap in the form of lack of standardized evaluation measures of explainability. Literature uses heterogeneous, generally subjective, measures of interpretability, making cross-model, cross-dataset, and cross-institution comparisons problematic.[7] There is a necessity to have validated and clinically based benchmarks that assess faithfulness, stability, and usefulness of explanations simultaneously.

Second, little prospective and real-world validation is a significant gap. The vast majority of XAI methods in the oncology field are retrospectively evaluated or in controlled settings and have limited information on their performance, strength, or impact on clinical decision-making in realistic workflows of various patient groups.[97]

Third, lack of sufficient integration of clinical context and domain knowledge is also a factor. Many of the explanations are feature-based or graphical and not patterned to oncologic logic, disease biology or cures. There is a need to conduct research to formulate concept-based and context-appropriate explanations that can be projected onto clinically meaningful constructs.[92]

Fourth, there is limited research on the gaps around human-AI interaction and usability. There is little empirical information on how clinicians process, believe or take action on XAI outputs, especially in high-stakes oncologic decision making.[118] End-user systematic studies are needed to maximize the explanation design, reduce automation bias, and enable successful human-AI interaction.

Fifth, there are still ethical, fairness, and regulatory vacuity. The detection of bias, the subgroup-specific reliability of the explanations, auditability, and the adherence to the regulatory rules are not covered consistently in the extant literature. Bringing the XAI development at the parity of developing emergent regulatory systems in medical AI systems of high risk is also a need that has not been met.

Lastly, causal and action gaps are still present. The majority of explanations are associative and not causal thus limiting their application in the treatment planning or intervention strategy. Causal inference, counterfactual reasoning, and quantification of uncertainty need to be combined in future studies that can help to improve the clinical utility of XAI in the medical field of oncology.[119]

Future Directions

The evenness between model complexity and interpretability is crucial for the advancement of AI applications in oncology. While advances in XAI have increased the transparency of machine learning models, several critical gaps must be addressed to ensure clinical relevance and widespread adoption.

(i) *Research Gaps in Real-World XAI Benchmarking*: One of the most pressing limitations is the lack of publicly available, high-quality, real-world datasets that include both clinical outcomes and ground-truth explanations. Most XAI evaluations are based on synthetic or retrospective datasets that do not reflect the dynamic nature of clinical decision-making. This limits the generalizability and practical validation of XAI systems in oncology.[120] There is a growing need for curated benchmark datasets that incorporate imaging, genomic, and clinical records, along with clinician-annotated explanation labels for comparative evaluation of XAI methods.

(ii) *Emerging XAI Methods for Complex Clinical Contexts*: Current XAI techniques like LIME and SHAP offer valuable insights, but their scope remains limited when applied to highly individualized, multimodal oncology models. Novel approaches such as contrastive explanations, which highlight why a prediction was made instead of an alternative diagnosis or treatment, and counterfactual explanations, which show what minimal changes would lead to a different outcome, are increasingly being explored.[121] These patient-specific explanation methods offer more actionable insights and align closely with how clinicians and patients reason about treatment options.

(iii) *Redefining Evaluation Metrics for XAI in Oncology*: Traditional performance metrics like accuracy, precision, and recall are insufficient for evaluating the utility of XAI in clinical practice. Future research should develop and standardize XAI-specific evaluation frameworks, including explanation fidelity, clinical usefulness, user satisfaction and trust, and comprehension and actionability, to assess the model's effectiveness in real-time decision-making. Integrating these metrics into AI development cycles will help ensure that explanations are not only technically correct but also practically valuable.

(iv) *Multimodal and Federated Future Models*: The future of oncology AI lies in multimodal data integration, where imaging, genomics, and structured clinical records are combined to generate holistic models. This will require more sophisticated XAI tools capable of disentangling and visualizing contributions from heterogeneous data sources.[86] Similarly, federated learning which enables training across decentralized datasets while preserving privacy will demand new XAI solutions that work consistently across sites, despite data heterogeneity. Additionally, continuous model validation will become essential as AI systems are deployed in clinical settings, requiring ongoing assessments to ensure that models remain accurate and relevant over time.

(v) *Human–AI collaboration*: One of the key future direction is the creation of human-AI collaborations where clarifications will enable effective communication between clinicians and models. The XAI has to become a complement to clinical judgment instead of its replacement, and facilitate shared decision-making, allow clinicians to question or challenge model outputs, and foster calibrated trust. Human-AI interaction also requires high-quality usability testing, formal training of the clinicians, and customized adaptive explanation interfaces based on the particular role of the user (such as oncologists, pathologists, and tumor board members).

(vi) *Ethical and regulatory systems*: The implementation of XAI in oncology should be informed by clear ethical values and regulatory frameworks that cover transparency, accountability, fairness, patient safety and data management. The proposed research directions should coincide with the clarification of the explainability approaches with new regulatory demands on high-risk AI systems, whereby explanations are auditable, reproducible, and medico-legally scrutinized. From regulatory standpoint, medical AI systems have dynamic oversight structures. The European Union propsed Artificial Intelligence Act defines AI systems used in the healthcare sector as a high-risk area and thus require strict adherence to transparency, human control, risk, and post-market surveillance.[122] Similarly, the regulatory authorities such as the U.S. Food and Drug Administration (FDA) emphasize the significance of good ML practice (GMLP), transparency of the models, performance monitoring, and change management of AI-based medical devices.[123] Although implicitly these regulatory frameworks require AI outputs to be sufficiently interpretable to support clinical accountability, enable auditability and guide the processes of decision-making, these regulatory frameworks do not dictate specific XAI methods. In the future, the development of XAI in the field of oncology needs to be more aligned with the changing regulatory and clinical demands. This alignment entails the development of explanation methodologies that are technically sound as well as

documentable, reproducible and auditable across the model iterations and deployment environments. Moreover, potential validation, ongoing performance checkup, and clear reporting of lack of certainty and failure modes will become the essential components of regulatory compliance. Moreover, explainability must be made to assist human-in-the-loop decision-making. The design makes clinicians be able to override, contextualize or query model outputs, thus avoiding passive adoption of algorithmic recommendations. All these factors reaffirm the need to have XAI solutions that are technically advanced and at the same time, more concrete in terms of regulatory and clinical practicality.

Limitations of the Study

While this study synthesizes the current state and potential of explainable AI in oncology, it primarily reflects perspectives from computer science, informatics, and biochemistry. This review has a number of limitations that should be considered keenly in drawing the implications of XAI to oncology. The absence of direct contributions from practicing clinical oncologists or pathologists is a notable limitation. To begin with, the research is a narrative synthesis, not a systematic review or meta-analysis; therefore, search strategy, study selection, risk-of-bias assessment were not performed according to the standardized guidelines, which could be the source of selection and publication bias. Second, the XAI literature is also not homogeneous in terms of algorithms, methods of explanation, datasets, endpoints, and reporting practices, thus making cross-study comparison challenging and preventing the possibility of making quantitative conclusions regarding clinical benefit. Third, much of the evidence is done on a retrospective basis or in a simulated environment with much less guiding prospective, multi-centered studies; therefore, actual performance, calibration, workflow integration, and patient-centered outcomes (eg, trust, shared decision-making, equity) are not fully characterized. Fourth, the explanations are frequently not validated with reference to clinician reasoning or plausible biological processes, and assessment is often based on proxy measures that are not necessarily relevant in practice in terms of utility or safety. Fifth, there is limited generalizability due to dataset shift, lack of representation of different populations and tumor subtypes, and confounding by imaging protocols and site effects, which can increase bias despite apparent accuracy. Lastly, explainability, privacy, and accountability policies put forward by the regulatory, legal, and ethical authorities are constantly changing, which means that some of the recommendations might require revision with the advancement of the standards. Future studies must focus on transparent reporting and reproducible evaluation both of model performance and of model explanations, do pre-registered prospective studies, multi-institutional validation and human-centered assessment of model explanations with clinicians and patients. These measures will be necessary to guarantee that XAI will improve interpretability and at the same time improve safety, fairness, and clinically meaningful outcomes in oncology.

Conclusion

XAI is a critical aspect of overcoming the long-standing black-box challenge that stands in the way of the reliable clinical application of AI in oncology. XAI methods are essential for enhancing the transparency, trust, and usability of AI systems in oncology. While widely used methods such as SHAP, LIME, and Grad-CAM can improve transparency, their clinical value depends on explanation faithfulness, stability, and usability, particularly in heterogeneous oncology settings spanning imaging, radiomics, genomics, and multimodal decision-making. This review attests that interpretability should be taken as a sociotechnical imperative, integrating technical elucidations with clinical workflow integration, human supervision and governance controls, and not be thought of as an algorithmic adjunct. To reach a broad coverage, this review goes beyond the analysis of SHAP, LIME, and Grad-CAM with intrinsically interpretable models; counterfactual and example-based explanations; concept-based and attention-based strategies, and surrogate and global interpretability strategies. The discussion highlights the fact that the usefulness of explanations in healthcare settings requires strict calibration, solid validation, and compliance with the clinical feasibility limits. By providing clear and interpretable model predictions, XAI techniques facilitate better decision-making, improve patient outcomes, and ensure that AI-driven treatments align with clinical expertise. However, overcoming challenges related to bias, data privacy, and regulatory compliance will require concerted efforts from researchers, clinicians, and policymakers. Addressing these gaps through prospective multi-center studies, uncertainty- and calibration-aware explanation interfaces, human-centered evaluation, and regulatory-ready reporting will be essential for advancing transparent, reliable, and clinically meaningful AI systems in oncology. As the number of people getting cancer is expected to rise around the world by 2050, solving the "black box" problem in oncology with XAI could lessen the bad effects of cancer by improving patient outcomes, should this terrible prediction come true.

Abbreviations

AI	Artificial Intelligence
CNN	Convolutional Neural Network
CT	Computed Tomography
CEM	Contrastive Explanation Method
EHR	Electronic Health Record
ELI5	Explain Like I'm 5
FDA	Food and Drug Administration
Grad-CAM	Gradient-weighted Class Activation Mapping
LIME	Local Interpretable Model-Agnostic Explanations
PDP-ICE	Partial Dependence Plot (PDP) and Individual Conditional Expectation (ICE)

PFI	Permutation Feature Importance
ProtoPNet	Prototypical Part Network
RCT	Randomized Controlled Trial
SHAP	Shapley Additive Explanations
XAI	Explainable Artificial Intelligence

Acknowledgements

None

ORCID iDs

Esther Ugo Alum https://orcid.org/0000-0003-4105-8615
Daniel Ejim Uti https://orcid.org/0000-0002-1129-1785

Ethics Approval

This study does not require an Ethics statement as it is based solely on publicly available data and does not involve direct interaction with human subjects.

Consent to Participate

Not applicable.

Consent to Publish Declaration

Not applicable.

Credit Authorship Contribution Statement

Conceptualization: EUA, DAE, BNA
Methodology: EUA, CKE, DAE, POE, VSM, JEE, BNA, DEU
Investigation: EUA, CKE, DAE, POE, VSM, JEE, BNA, DEU
Data Interpretation: EUA, CKE, DAE, POE, VSM, JEE, BNA, DEU
Resources: EUA, CKE, DAE, POE, VSM, JEE, BNA, DEU
Supervision: VSM, JEE, CKE
Validation: POE, BNA
Visualization: EUA, VSM
Writing – original draft: EUA, DEU
Writing – review & editing: EUA, CKE, DAE, POE, VSM, JEE, BNA, DEU

Funding

The authors received no financial support for the research, authorship, and/or publication of this article.

Declaration of Conflicting Interests

The authors declared no potential conflicts of interest with respect to the research, authorship, and/or publication of this article.

Competing Interest Declaration

We declare no competing interests.

Availability Statement

All data are utilized in the manuscript.

Clinical Trial Date of Registration

Not applicable.

Clinical Trial Registration Number

Not applicable.

Clinical Trial Registry

Not applicable.

AI Disclosure Statement

During the preparation of this manuscript, the authors used Quilbot and Grammarly tools for language editing, and grammar correction improvement. However, after its use, the authors thoroughly reviewed, verified, and revised all assisted content to ensure accuracy and originality. The authors, therefore, take full responsibility for the integrity and final content of the published article.

Supplemental Material

Supplemental material for this article is available online.

References

1. Veziant J, Villéger R, Barnich N, Bonnet M. Gut microbiota as potential biomarker and/or therapeutic target to improve the management of cancer: focus on colibactin-producing escherichia coli in colorectal cancer. *Cancers (Basel)*. 2021;13(9):2215. doi:10.3390/cancers13092215
2. Alum EU. AI-driven biomarker discovery: enhancing precision in cancer diagnosis and prognosis. *Discov Onc*. 2025;16(1):313. doi:10.1007/s12672-025-02064-7
3. Alowais SA, Alghamdi SS, Alsuhebany N, et al. Revolutionizing healthcare: the role of artificial intelligence in clinical practice. *BMC Med Educ.* 2023;23(1):689. doi:10.1186/s12909-023-04698-z
4. Vora LK, Gholap AD, Jetha K, Thakur RRS, Solanki HK, Chavda VP. Artificial intelligence in pharmaceutical technology and drug delivery design. *Pharmaceutics*. 2023;15(7):1916. doi:10.3390/pharmaceutics15071916
5. Ahmed MI, Spooner B, Isherwood J, Lane M, Orrock E, Dennison A. A systematic review of the barriers to the implementation of artificial intelligence in healthcare. *Cureus*. 15:e46454. doi:10.7759/cureus.46454
6. Fehr J, Citro B, Malpani R, Lippert C, Madai VI. A trustworthy AI reality-check: the lack of transparency of artificial intelligence products in healthcare. *Front Digit Health*. 2024;6:1267290. doi:10.3389/fdgth.2024.1267290
7. Sadeghi Z, Alizadehsani R, Cifci MA, et al. A review of explainable artificial intelligence in healthcare. *Comput Electr Eng.* 2024;118:109370. doi:10.1016/j.compeleceng.2024.109370
8. Cui S, Traverso A, Niraula D, et al. Interpretable artificial intelligence in radiology and radiation oncology. *Br J Radiol*. 2023;96(1150):20230142. doi:10.1259/bjr.20230142
9. Du Z, Yang L, He H, Wu X, Qi X, Zhang Y. Artificial intelligence for the noninvasive diagnosis of clinically significant portal hypertension. *EngMedicine*. 2025;2(2):100069. doi:10.1016/j.engmed.2025.100069
10. Vivek Khanna V, Chadaga K, Sampathila N, et al. Explainable artificial intelligence-driven gestational diabetes mellitus

prediction using clinical and laboratory markers. *Cogent Eng*. 2024;11(1):2330266. doi:10.1080/23311916.2024.2330266
11. Goswami NG, Goswami A, Sampathila N, Bairy MG, Chadaga K, Belurkar S. Detection of sickle cell disease using deep neural networks and explainable artificial intelligence. *J Intell Syst.* 2024;33(1). doi:10.1515/jisys-2023-0179
12. Collins GS, Chester-Jones M, Gerry S, et al. Clinical prediction models using machine learning in oncology: Challenges and recommendations. *BMJ Oncol*. 2025;4(1):e000914. doi:10.1136/bmjonc-2025-000914
13. Saarela M, Podgorelec V. Recent applications of explainable AI (XAI): a systematic literature review. *Appl Sci.* 2024;14(19):8884. doi:10.3390/app14198884
14. Salih AM, Raisi-Estabragh Z, Galazzo IB, et al. A perspective on explainable artificial intelligence methods: SHAP and LIME. *Adv Intell Syst*. 2025;7(1):2400304. doi:10.1002/aisy.202400304
15. Ranjbaran G, Recupero DR, Roy CK, Schneider KA. C-SHAP: a hybrid method for fast and efficient interpretability. *Appl Sci.* 2025;15(2):672. doi:10.3390/app15020672
16. Ponce-Bobadilla AV, Schmitt V, Maier CS, Mensing S, Stodtmann S. Practical guide to SHAP analysis: explaining supervised machine learning model predictions in drug development. *Clin Transl Sci*. 2024;17(11):e70056. doi:10.1111/cts.70056
17. Babaei G, Giudici P. InstanceSHAP: an instance-based estimation approach for Shapley values. *Behaviormetrika*. 2024;51(1):425-439. doi:10.1007/s41237-023-00208-z
18. Aas K, Jullum M, Løland A. Explaining individual predictions when features are dependent: more accurate approximations to Shapley values. *Artif Intell.* 2021;298:103502. doi:10.1016/j.artint.2021.103502
19. Talaat FM, Gamel SA, El-Balka RM, Shehata M, ZainEldin H. Grad-CAM enabled breast cancer classification with a 3D inception-ResNet V2: empowering radiologists with explainable insights. *Cancers (Basel)*. 2024;16(21):3668. doi:10.3390/cancers16213668
20. Wang F, Li Y, Zeng T. Deep learning of radiology-genomics integration for computational oncology: a mini review. *Comput Struct Biotechnol J.* 2024;23:2708-2716. doi:10.1016/j.csbj.2024.06.019
21. Carles-Bou JL, Carmona EJ. Achieving faithful explainability in feedforward neural networks through accurately computed feature attribution. *Neural Netw.* 2026;195:108277. doi:10.1016/j.neunet.2025.108277
22. Jha A, Aicher JK, Gazzara MR, Singh D, Barash Y. Enhanced integrated gradients: improving interpretability of deep learning models using splicing codes as a case study. *Genome Biol.* 2020;21(1):149. doi:10.1186/s13059-020-02055-7
23. Sigut J, Fumero F, Arnay R, Estévez J, Díaz-Alemán T. Interpretable surrogate models to approximate the predictions of convolutional neural networks in glaucoma diagnosis. *Mach Learn Sci Technol*. 2023;4(4):045024. doi:10.1088/2632-2153/ad0798
24. Najjar R. Redefining radiology: a review of artificial intelligence integration in medical imaging. *Diagnostics (Basel)*. 2023;13(17):2760. doi:10.3390/diagnostics13172760
25. Zhang Y-P, Zhang X-Y, Cheng Y-T, et al. Artificial intelligence-driven radiomics study in cancer: the role of feature engineering and modeling. *Mil Med Res.* 2023;10(1):22. doi:10.1186/s40779-023-00458-8
26. Alkhanbouli R, Matar Abdulla Almadhaani H, Alhosani F, Simsekler MCE. The role of explainable artificial intelligence in disease prediction: a systematic literature review and future research directions. *BMC Med Inform Decis Mak*. 2025;25(1):110. doi:10.1186/s12911-025-02944-6
27. Ribeiro MT, Singh S, Guestrin C. "Why should I trust you?": explaining the predictions of any classifier. In: Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining. Association for Computing Machinery, New York, NY, USA; 2016:1135-1144.
28. Lundberg SM, Lee S-I. A unified approach to interpreting model predictions. In: Proceedings of the 31st International Conference on Neural Information Processing Systems. Curran Associates Inc., Red Hook, NY, USA; 2017:4768-4777.
29. Lundberg SM, Erion G, Chen H, et al. From local explanations to global understanding with explainable AI for trees. *Nat Mach Intell*. 2020;2(1):56-67. doi:10.1038/s42256-019-0138-9
30. Selvaraju RR, Cogswell M, Das A, Vedantam R, Parikh D, Batra D. Grad-CAM: visual explanations from deep networks via gradient-based localization. In: 2017 IEEE International Conference on Computer Vision (ICCV); 2017:618-626.
31. Zhou B, Khosla A, Lapedriza A, Oliva A, Torralba A. Learning deep features for discriminative localization. In: 2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR). IEEE, Las Vegas, NV, USA; 2016:2921-2929.
32. Sundararajan M, Taly A, Yan Q. Axiomatic attribution for deep networks. In: Proceedings of the 34th International Conference on Machine Learning. PMLR; 2017:3319-3328.
33. Shrikumar A, Greenside P, Kundaje A. Learning important features through propagating activation differences. In: Proceedings of the 34th International Conference on Machine Learning. PMLR; 2017:3145-3153.
34. Rizzo M, Veneri A, Marcuzzo M, et al. Machine learning models explanations as interpretations of evidence: a theoretical framework of explainability and its implications on high-stakes biomedical decision-making. *BMC Med Res Methodol*. 2025;25(S1):282. doi:10.1186/s12874-025-02703-1
35. Lu S-C, Swisher CL, Chung C, Jaffray D, Sidey-Gibbons C. On the importance of interpretable machine learning predictions to inform clinical decision making in oncology. *Front Oncol*. 2023;13:1129380. doi:10.3389/fonc.2023.1129380
36. Jia Z, Zeng X, Duan H, Lu X, Li H. A patient–similarity–based model for diagnostic prediction. *Int J Med Inform*. 2020 Mar;135:104073. doi: 10.1016/j.ijmedinf.2019.104073. Epub 2019 Dec 30. Erratum in: *Int J Med Inform*. 2020 May;137:104100. doi: 10.1016/j.ijmedinf.2020.104100. PMID: 31923816.
37. Hanna MG, Pantanowitz L, Jackson B, et al. Ethical and bias considerations in artificial intelligence/machine learning. *Mod Pathol.* 2025;38(3):100686. doi:10.1016/j.modpat.2024.100686
38. Andringa J, Baptista ML, Santos BF. Counterfactual explanations for remaining useful life estimation within a Bayesian

framework. *Inf Fusion.* 2025;118:102972. doi:10.1016/j.inffus.2025.102972
39. Dickerman BA, Hernán MA. Counterfactual prediction is not only for causal inference. *Eur J Epidemiol*. 2020;35(7):615-617. doi:10.1007/s10654-020-00659-8
40. Vats A, Pedersen M, Mohammed A. Concept-based reasoning in medical imaging. *Int J Comput Assist Radiol Surg*. 2023;18(7):1335-1339. doi:10.1007/s11548-023-02920-3
41. An J, Joe I. Attention map-guided visual explanations for deep neural networks. *Appl Sci.* 2022;12(8):3846. doi:10.3390/app12083846
42. Salvi M, Seoni S, Campagner A, et al. Explainability and uncertainty: two sides of the same coin for enhancing the interpretability of deep learning models in healthcare. *Int J Med Inf.* 2025;197:105846. doi:10.1016/j.ijmedinf.2025.105846
43. Rawal A, Raglin A, Rawat DB, Sadler BM, McCoy J. Causality for trustworthy artificial intelligence: status, challenges and perspectives. *ACM Comput Surv*. 2025;57(6):146, 1-30. doi:10.1145/3665494.
44. Mehdiyev N, Majlatow M, Fettke P. Integrating permutation feature importance with conformal prediction for robust explainable artificial intelligence in predictive process monitoring. *Eng Appl Artif Intell.* 2025;149:110363. doi:10.1016/j.engappai.2025.110363
45. Cox LA. What is an exposure-response curve? *Glob Epidemiol*. 2023;6:100114. doi:10.1016/j.gloepi.2023.100114
46. Baniecki H, Parzych D, Biecek P. The grammar of interactive explanatory model analysis. *Data Min Knowl Discov*. 2024;38(5):2596-2632. doi:10.1007/s10618-023-00924-w
47. Shirasawa R, Takaki K, Miyao T. Generalizability improvement of interpretable symbolic regression models for quantitative structure–activity relationships. *ACS Omega*. 2024;9(8):9463-9474. doi:10.1021/acsomega.3c09047
48. Ennab M, Mcheick H. Advancing AI interpretability in medical imaging: a comparative analysis of pixel-level interpretability and Grad-CAM models. *Mach Learn Knowl Extr*. 2025;7(1):12. doi:10.3390/make7010012
49. Ghasemi A, Hashtarkhani S, Schwartz DL, Shaban-Nejad A. Explainable artificial intelligence in breast cancer detection and risk prediction: a systematic scoping review. *Cancer Innov*. 2024;3(5):e136. doi:10.1002/cai2.136
50. Hassan SU, Abdulkadir SJ, Zahid MSM, Al-Selwi SM. Local interpretable model-agnostic explanation approach for medical imaging analysis: a systematic literature review. *Comput Biol Med.* 2025;185:109569. doi:10.1016/j.compbiomed.2024.109569
51. Rohrer JM. Thinking clearly about correlations and causation: graphical causal models for observational data. *Adv Methods Pract Psychol Sci*. 2018;1(1):27-42. doi:10.1177/2515245917745629
52. Sarker IH. Deep learning: a comprehensive overview on techniques, taxonomy, applications and research directions. *SN Comput Sci*. 2021;2(6):420. doi:10.1007/s42979-021-00815-1
53. Zhong T, Zhang Q, Huang J, Wu M, Ma S. Heterogeneity analysis via integrating multi-sources high-dimensional data with applications to cancer studies. *Stat Sin*. 2023;33:729-758. doi:10.5705/ss.202021.0002
54. Jiménez-Sánchez A, Avlona N-R, de Boer S, et al. In the picture: medical imaging datasets, artifacts, and their living review. In: Proceedings of the 2025 ACM Conference on Fairness, Accountability, and Transparency. Association for Computing Machinery, New York, NY, USA; 2025:511-531.
55. Lee S, Yin C, Zhang P. Stable clinical risk prediction against distribution shift in electronic health records. *Patterns (N Y)*. 2023;4(9):100828. doi:10.1016/j.patter.2023.100828
56. Walraven JEW, van der Hel OL, van der Hoeven JJM, Lemmens VEPP, Verhoeven RHA, Desar IME. Factors influencing the quality and functioning of oncological multidisciplinary team meetings: results of a systematic review. *BMC Health Serv Res.* 2022;22(1):829. doi:10.1186/s12913-022-08112-0
57. Hagenbuchner M. The black box problem of AI in oncology. *J Phys Conf Ser*. 2020;1662(1):012012. doi:10.1088/1742-6596/1662/1/012012
58. Ladbury C, Zarinshenas R, Semwal H, et al. Utilization of model-agnostic explainable artificial intelligence frameworks in oncology: a narrative review. *Transl Cancer Res.* 2022;11(10):3853-3868. doi:10.21037/tcr-22-1626
59. Sathyan A, Weinberg AI, Cohen K. Interpretable AI for bio-medical applications. *Com Eng Sys*. 2022;2(4):18. doi:10.20517/ces.2022.41
60. Vimbi V, Shaffi N, Mahmud M. Interpreting artificial intelligence models: a systematic review on the application of LIME and SHAP in Alzheimer's disease detection. *Brain Inform.* 2024;11(1):10. doi:10.1186/s40708-024-00222-1
61. London AJ. Artificial intelligence and black-box medical decisions: accuracy versus explainability. *Hastings Cent Rep*. 2019;49(1):15-21. doi:10.1002/hast.973
62. Ugwu OP-C, Alum EU, Ugwu JN, et al. Harnessing technology for infectious disease response in conflict zones: challenges, innovations, and policy implications. *Medicine (Baltimore)*. 2024;103(28):e38834. doi:10.1097/MD.0000000000038834
63. Taye MM. Understanding of machine learning with deep learning: architectures, workflow, applications and future directions. *Computers*. 2023;12(5):91. doi:10.3390/computers12050091
64. Yang G, Ye Q, Xia J. Unbox the black-box for the medical explainable AI via multi-modal and multi-centre data fusion: a mini-review, two showcases and beyond. *Inf Fusion.* 2022;77:29-52. doi:10.1016/j.inffus.2021.07.016
65. Singh G, Kamalja A, Patil R, Karwa A, Tripathi A, Chavan P. A comprehensive assessment of artificial intelligence applications for cancer diagnosis. *Artif Intell Rev*. 2024;57(7):179. doi:10.1007/s10462-024-10783-6
66. Bera K, Schalper KA, Rimm DL, Velcheti V, Madabhushi A. Artificial intelligence in digital pathology — new tools for diagnosis and precision oncology. *Nat Rev Clin Oncol*. 2019;16(11):703-715. doi:10.1038/s41571-019-0252-y
67. O'Connor O, McVeigh TP. Increasing use of artificial intelligence in genomic medicine for cancer care- the promise and potential pitfalls. *BJC Rep*. 2025;3(1):20. doi:10.1038/s44276-025-00135-4
68. Perniciano A, Loddo A, Di Ruberto C, Pes B. Insights into radiomics: impact of feature selection and classification.

Multimed Tools Appl. 2025;84(26):31695-31721. doi:10.1007/s11042-024-20388-4
69. Prentzas N, Kakas A, Pattichis CS. Explainable AI applications in the Medical Domain: a systematic review; 2023. http://arxiv.org/abs/2308.05411
70. Rane N, Choudhary S, Rane J. Explainable Artificial Intelligence (XAI) in healthcare: interpretable models for clinical decision support; 2023. https://papers.ssrn.com/abstract=4637897
71. Zhang Y, Weng Y, Lund J. Applications of explainable artificial intelligence in diagnosis and surgery. *Diagnostics*. 2022;12(2):237. doi:10.3390/diagnostics12020237
72. El Shawi R, Sherif Y, Al-Mallah M, Sakr S. Interpretability in HealthCare a comparative study of local machine learning interpretability techniques. In: 2019 IEEE 32nd International Symposium on Computer-Based Medical Systems (CBMS); 2019:275-280.
73. Pertuz S, Ortega D, Suarez É, et al. Saliency of breast lesions in breast cancer detection using artificial intelligence. *Sci Rep*. 2023;13(1):20545. doi:10.1038/s41598-023-46921-3
74. Hassan B, Raja H, Hassan T, et al. A comprehensive review of artificial intelligence models for screening major retinal diseases. *Artif Intell Rev*. 2024;57(5):111. doi:10.1007/s10462-024-10736-z
75. Retzlaff CO, Angerschmid A, Saranti A, et al. Post-hoc vs ante-hoc explanations: XAI design guidelines for data scientists. *Cogn Syst Res.* 2024;86:101243. doi:10.1016/j.cogsys.2024.101243
76. Semba RD, Askari S, Gibson S, Bloem MW, Kraemer K. The potential impact of climate change on the micronutrient-rich food supply. *Adv Nutr*. 2021;13(1):80-100. doi:10.1093/advances/nmab104
77. Díaz-Rodríguez N, Del Ser J, Coeckelbergh M, López de Prado M, Herrera-Viedma E, Herrera F. Connecting the dots in trustworthy artificial intelligence: from AI principles, ethics, and key requirements to responsible AI systems and regulation. *Inf Fusion.* 2023;99:101896. doi:10.1016/j.inffus.2023.101896
78. Aldoseri A, Al-Khalifa KN, Hamouda AM. Re-thinking data strategy and integration for artificial intelligence: concepts, opportunities, and challenges. *Appl Sci.* 2023;13(12):7082. doi:10.3390/app13127082
79. Yelne S, Chaudhary M, Dod K, Sayyad A, Sharma R. Harnessing the power of AI: a comprehensive review of its impact and challenges in nursing science and healthcare. *Cureus*. 15:e49252. doi:10.7759/cureus.49252
80. Chow R, Midroni J, Kaur J, et al. Use of artificial intelligence for cancer clinical trial enrollment: a systematic review and meta-analysis. *J Natl Cancer Inst*. 2023;115(4):365-374. doi:10.1093/jnci/djad013
81. Goto S, Ozawa H. The importance of external validation for neural network models. *JACC Adv*. 2023;2(8):100610. doi:10.1016/j.jacadv.2023.100610
82. Khosravi M, Zare Z, Mojtabaeian SM, Izadi R. Artificial intelligence and decision-making in healthcare: a thematic analysis of a systematic review of reviews. *Health Serv Res Manag Epidemiol*. 2024;11:23333928241234863. doi:10.1177/23333928241234863
83. Zhang J, Zhang Z. Ethics and governance of trustworthy medical artificial intelligence. *BMC Med Inform Decis Mak.* 2023;23(1):7. doi:10.1186/s12911-023-02103-9
84. Holscher HD. Dietary fiber and prebiotics and the gastrointestinal microbiota. *Gut Microbes*. 2017;8(2):172-184. doi:10.1080/19490976.2017.1290756
85. van Leersum CM, Maathuis C. Human centred explainable AI decision-making in healthcare. *J Responsible Technol*. 2025;21:100108. doi:10.1016/j.jrt.2025.100108
86. Brockmueller A, Buhrmann C, Moravejolahkami AR, Shakibaei M. Resveratrol and p53: how are they involved in CRC plasticity and apoptosis? *J Adv Res.* 2024;66:181-195. doi:10.1016/j.jare.2024.01.005
87. Butow P, Hoque E. Using artificial intelligence to analyse and teach communication in healthcare. *Breast*. 2020;50:49-55. doi:10.1016/j.breast.2020.01.008
88. Carcagnì P, Leo M, Del Coco M, Distante C, De Salve A. Convolution neural networks and self-attention learners for Alzheimer dementia diagnosis from brain MRI. *Sensors*. 2023;23(3):1694. doi:10.3390/s23031694
89. Alum EU, Ugwu OP-C. Artificial intelligence in personalized medicine: transforming diagnosis and treatment. *Discov Appl Sci*. 2025;7(3):193. doi:10.1007/s42452-025-06625-x
90. Mohammed S, Malhotra N. Ethical and regulatory challenges in machine learning-based healthcare systems: a review of implementation barriers and future directions. *BenchCouncil Trans Benchmarks Stand Eval*. 2025;5(1):100215. doi:10.1016/j.tbench.2025.100215
91. Longoni C, Bonezzi A, Morewedge CK. Resistance to medical artificial intelligence. *J Consum Res.* 2019;46(4):629-650. doi:10.1093/jcr/ucz013
92. Abbas Q, Jeong W, Lee SW. Explainable AI in clinical decision support systems: a meta-analysis of methods, applications, and usability challenges. *Healthcare*. 2025;13(17):2154. doi:10.3390/healthcare13172154
93. Gulum MA, Trombley CM, Kantardzic M. A review of explainable deep learning cancer detection models in medical imaging. *Appl Sci.* 2021;11(10):4573. doi:10.3390/app11104573
94. Givisis I, Kalatzis D, Christakis C, Kiouvrekis Y. Comparing explainable AI models: SHAP, LIME, and their role in electric field strength prediction over urban areas. *Electronics (Basel).* 2025;14(23):4766. doi:10.3390/electronics14234766
95. Crispin-Ortuzar M, Woitek R, Reinius MAV, et al. Integrated radiogenomics models predict response to neoadjuvant chemotherapy in high grade serous ovarian cancer. *Nat Commun.* 2023;14(1):6756. doi:10.1038/s41467-023-41820-7
96. Lamane H, Mouhir L, Moussadek R, Baghdad B, Kisi O, El Bilali A. Interpreting machine learning models based on SHAP values in predicting suspended sediment concentration. *Int J Sediment Res.* 2025;40:91-107. doi:10.1016/j.ijsrc.2024.10.002
97. Shifa N, Saleh M, Akbari Y, Al Maadeed S. A review of explainable AI techniques and their evaluation in mammography for breast cancer screening. *Clin Imaging.* 2025;123:110492. doi:10.1016/j.clinimag.2025.110492
98. Houssein EH, Gamal AM, Younis EMG, Mohamed E. Explainable artificial intelligence for medical imaging systems

using deep learning: a comprehensive review. *Cluster Comput.* 2025;28(7):469. doi:10.1007/s10586-025-05281-5

99. Starcke J, Spadafora J, Spadafora J, Spadafora P, Toma M. The effect of data leakage and feature selection on machine learning performance for early Parkinson's disease detection. *Bioengineering (Basel)*. 2025;12(8):845. doi:10.3390/bioengineering12080845
100. Reinhammar R, Waernbaum I. Covariate selection strategies and estimands - a review of current practice of risk factor analysis from a causal perspective. *BMC Med Res Methodol*. 2025;25(1):260. doi:10.1186/s12874-025-02704-0
101. Norton K-A, Bergman D, Jain HV, Jackson T. Advances in surrogate modeling for biological agent-based simulations: trends, challenges, and future prospects. *J Math Biol.* 2025;92(1):6. doi:10.1007/s00285-025-02318-6
102. Van Calster B, McLernon DJ, van Smeden M, Wynants L, Steyerberg EW. Calibration: the Achilles heel of predictive analytics. *BMC Med.* 2019;17(1):230. doi:10.1186/s12916-019-1466-7
103. Pálfi B, Arora K, Prociuk D, Kostopoulou O. Risk prediction algorithms and clinical judgment: impact of advice distance, social proof, and feature-importance explanations. *Comput Human Behav.* 2024;153:108102. doi:10.1016/j.chb.2023.108102
104. Benčević M, Habijan M, Galić I, Babin D, Pižurica A. Understanding skin color bias in deep learning-based skin lesion segmentation. *Comput Methods Programs Biomed.* 2024;245:108044. doi:10.1016/j.cmpb.2024.108044
105. Khor S, Haupt EC, Hahn EE, Lyons LJL, Shankaran V, Bansal A. Racial and ethnic bias in risk prediction models for colorectal cancer recurrence when race and ethnicity are omitted as predictors. *JAMA Netw Open*. 2023;6(6):e2318495. doi:10.1001/jamanetworkopen.2023.18495
106. Gu T, Pan W, Yu J, et al. Mitigating bias in AI mortality predictions for minority populations: a transfer learning approach. *BMC Med Inform Decis Mak.* 2025;25(1):30. doi:10.1186/s12911-025-02862-7
107. Mienye ID, Obaido G, Jere N, et al. A survey of explainable artificial intelligence in healthcare: concepts, applications, and challenges. *Inform Med Unlocked*. 2024;51:101587. doi:10.1016/j.imu.2024.101587
108. Zhu B, Niu L. A privacy-preserving federated learning scheme with homomorphic encryption and edge computing. *Alexandria Eng J.* 2025;118:11-20. doi:10.1016/j.aej.2024.12.070
109. Yurdem B, Kuzlu M, Gullu MK, Catak FO, Tabassum M. Federated learning: OVERVIEW, strategies, applications, tools and future directions. *Heliyon*. 2024;10(19):e38137. doi:10.1016/j.heliyon.2024.e38137
110. Ebad SA, Alhashmi A, Amara M, Miled AB, Saqib M. Artificial intelligence-based software as a medical device (AI-SaMD): a systematic review. *Healthcare*. 2025;13(7):817. doi:10.3390/healthcare13070817
111. Longo L, Brcic M, Cabitza F, et al. Explainable artificial intelligence (XAI) 2.0: a manifesto of open challenges and interdisciplinary research directions. *Inf Fusion.* 2024;106:102301. doi:10.1016/j.inffus.2024.102301
112. Goel I, Bhaskar Y, Kumar N, et al. Role of AI in empowering and redefining the oncology care landscape: perspective from a developing nation. *Front Digit Health*. 2025;7. doi:10.3389/fdgth.2025.1550407
113. Noor AA, Manzoor A, Mazhar Qureshi MD, Qureshi MA, Rashwan W. Unveiling explainable AI in healthcare: current trends, challenges, and future directions. *WIREs Data Min Knowl Discov*. 2025;15(2):e70018. doi:10.1002/widm.70018
114. Cappelleri JC, Bushmakin AG. Measurement of patient-reported outcomes of health services. In: *Methods in Health Services Research*. Springer; 2017:1-21.
115. Abgrall G, Holder AL, Chelly Dagdia Z, Zeitouni K, Monnet X. Should AI models be explainable to clinicians? *Crit Care.* 2024;28(1):301. doi:10.1186/s13054-024-05005-y
116. Fernainy P, Cohen AA, Murray E, Losina E, Lamontagne F, Sourial N. Rethinking the pros and cons of randomized controlled trials and observational studies in the era of big data and advanced methods: a panel discussion. *BMC Proc.* 2024;18(S2):1. doi:10.1186/s12919-023-00285-8
117. Jeyaraman M, Balaji S, Jeyaraman N, Yadav S. Unraveling the ethical enigma: artificial intelligence in healthcare. *Cureus*. 15:e43262. doi:10.7759/cureus.43262
118. Subramanian HV, Canfield C, Shank DB. Designing explainable AI to improve human-AI team performance: a medical stakeholder-driven scoping review. *Artif Intell Med.* 2024;149:102780. doi:10.1016/j.artmed.2024.102780
119. Zahoor S, Liò P, Dias G, Hasanuzzaman M. Integrating probabilistic trees and causal networks for clinical and epidemiological data. *Artif Intell Med.* 2026;173:103350. doi:10.1016/j.artmed.2026.103350
120. Giuffrè M, Shung DL. Harnessing the power of synthetic data in healthcare: innovation, application, and privacy. *NPJ Digit Med*. 2023;6(1):186. doi:10.1038/s41746-023-00927-3
121. Mertes S, Huber T, Weitz K, Heimerl A, André E. GANterfactual—counterfactual explanations for medical non-experts using generative adversarial learning. *Front Artif Intell*. 2022;5. doi:10.3389/frai.2022.825565
122. Busch F, Kather JN, Johner C, et al. Navigating the European Union Artificial Intelligence Act for healthcare. *NPJ Digit Med*. 2024;7(1):210. doi:10.1038/s41746-024-01213-6
123. Singh V, Cheng S, Kwan AC, Ebinger J. United States Food and Drug Administration regulation of clinical software in the era of artificial intelligence and machine learning. *Mayo Clin Proc Digit Health*. 2025;3(3):100231. doi:10.1016/j.mcpdig.2025.100231