
Citation:

Alam, KN and Zadeh, PB and Sheikh-Akbari, A (2025) Attribution-Based Explainability in Medical Imaging: A Critical Review on Explainable Computer Vision (X-CV) Techniques and Their Applications in Medical AI. *Electronics*, 14 (15). pp. 1-26. ISSN 2079-9292 DOI: <https://doi.org/10.3390/electronics14153024>

Link to Leeds Beckett Repository record:

<https://eprints.leedsbeckett.ac.uk/id/eprint/12291/>

Document Version:

Article (Published Version)

Creative Commons: Attribution 4.0

© 2025 by the authors

The aim of the Leeds Beckett Repository is to provide open access to our research, as required by funder policies and permitted by publishers and copyright law.




The Leeds Beckett repository holds a wide range of publications, each of which has been checked for copyright and the relevant embargo period has been applied by the Research Services team.

We operate on a standard take-down policy. If you are the author or publisher of an output and you would like it removed from the repository, please [contact us](#) and we will investigate on a case-by-case basis.

Each thesis in the repository has been cleared where necessary by the author for third party copyright. If you would like a thesis to be removed from the repository or believe there is an issue with copyright, please contact us on openaccess@leedsbeckett.ac.uk and we will investigate on a case-by-case basis.

Review

Attribution-Based Explainability in Medical Imaging: A Critical Review on Explainable Computer Vision (X-CV) Techniques and Their Applications in Medical AI

Kazi Nabiul Alam , Pooneh Bagheri Zadeh  and Akbar Sheikh-Akbari * 

School of Built Environment, Engineering and Computing, Leeds Beckett University, Leeds LS6 3QS, UK; k.alam3742@student.leedsbeckett.ac.uk (K.N.A.); p.bagheri-zadeh@leedsbeckett.ac.uk (P.B.Z.)

* Correspondence: a.sheikh-akbari@leedsbeckett.ac.uk (A.S.-A.)

Abstract

One of the largest future applications of computer vision is in the healthcare industry. Computer vision tasks are generally implemented in diverse medical imaging scenarios, including detecting or classifying diseases, predicting potential disease progression, analyzing cancer data for advancing future research, and conducting genetic analysis for personalized medicine. However, a critical drawback of using Computer Vision (CV) approaches is their limited reliability and transparency. Clinicians and patients must comprehend the rationale behind predictions or results to ensure trust and ethical deployment in clinical settings. This demonstrates the adoption of the idea of Explainable Computer Vision (X-CV), which enhances vision-relative interpretability. Among various methodologies, attribution-based approaches are widely employed by researchers to explain medical imaging outputs by identifying influential features. This article solely aims to explore how attribution-based X-CV methods work in medical imaging, what they are good for in real-world use, and what their main limitations are. This study evaluates X-CV techniques by conducting a thorough review of relevant reports, peer-reviewed journals, and methodological approaches to obtain an adequate understanding of attribution-based approaches. It explores how these techniques tackle computational complexity issues, improve diagnostic accuracy and aid clinical decision-making processes. This article intends to present a path that generalizes the concept of trustworthiness towards AI-based healthcare solutions.

Keywords: attribution; XCV; artificial intelligence; explainability; medical imaging



Academic Editors: Jinwen Liang and Jixin Zhang

Received: 30 May 2025

Revised: 30 June 2025

Accepted: 9 July 2025

Published: 29 July 2025

Citation: Alam, K.N.; Zadeh, P.B.; Sheikh-Akbari, A. Attribution-Based Explainability in Medical Imaging: A Critical Review on Explainable Computer Vision (X-CV) Techniques and Their Applications in Medical AI. *Electronics* **2025**, *14*, 3024. <https://doi.org/10.3390/electronics14153024>

Copyright: © 2025 by the authors. Licensee MDPI, Basel, Switzerland. This article is an open access article distributed under the terms and conditions of the Creative Commons Attribution (CC BY) license (<https://creativecommons.org/licenses/by/4.0/>).

1. Introduction

Medical imaging benefits from artificial intelligence (AI) and deep-learning technology, which helps in diagnosis, treatment planning and prediction modeling. Through the integration of deep learning and AI into medical imaging diagnostics, the accuracy of diagnosis has increased while the efficiency improved, leading to earlier detection and individualized treatment approaches. Convolutional Neural Networks (CNNs) represent deep learning's foundation for medical image pattern extraction which exceeds traditional methods in tumour detection and fracture and cardiovascular abnormality identification within X-rays and MRIs and CT scans. Recent research shows that CNNs perform similarly to radiologists in two applications: breast cancer detection from mammograms [1] and pneumonia identification from chest X-rays [2]. The implementation of AI tools in healthcare facilities enables the automation of complex image processes and reduces radiologist fatigue, and

speeds up [3] diagnosis through workflow optimization. The adoption of AI technology remains challenging because data quality issues and model interpretability problems and generalizability challenges across different patient demographics represent [4] essential obstacles. The ongoing research and adoption of AI systems to enhance clinical decision-making continues because of their potential benefits. Generative adversarial networks and transformer architectures have enabled progress in solving rare disease data scarcity problems which enhances image reconstruction techniques and noise reduction, and produces training images [5]. Artificial intelligence models show great promise in forecasting disease development through analysis [6] of longitudinal imaging information such as brain MRIs for Alzheimer’s disease staging. The analysis of medical imaging through AI becomes more detailed by integrating electronic health records with multi-modal data which enhances [7] both risk assessment and treatment effectiveness. Medical imaging AI deployment requires strong ethical guidelines [8] that address algorithmic bias and protect patient privacy. The complete benefits of artificial intelligence for medical imaging will become accessible by fostering joint work between doctors and engineers and legislators throughout the development of these technologies.

Figure 1 refers to the procedure of explanation-based vision model’s help in medical image analysis based in the healthcare sector.

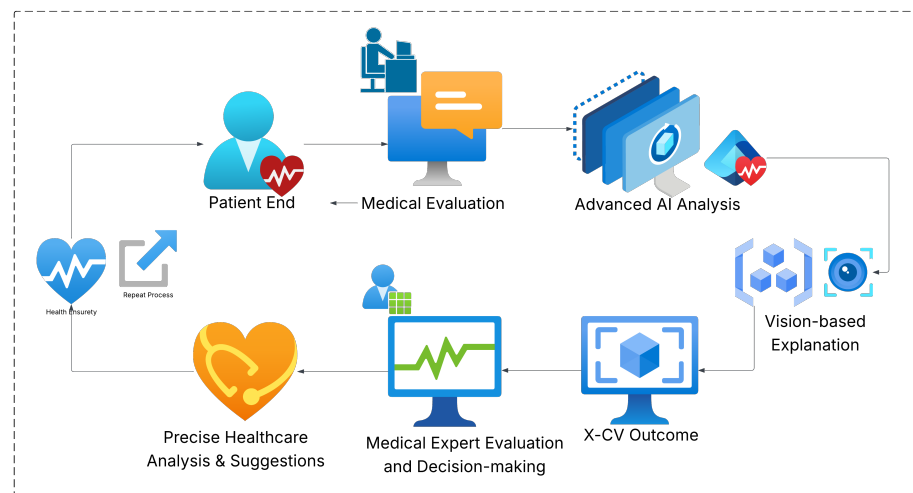


Figure 1. X-CV procedure: a guide to how the XCV process works in healthcare.

The rapid adoption of artificial intelligence (AI) in medical applications has spotlighted the “black box” nature of many deep learning models, particularly in high-stakes domains like diagnostics and treatment planning. These models, often based on complex neural networks, deliver impressive predictive accuracy but lack transparency in how they arrive at decisions, raising concerns [9] about reliability and accountability. Explainability in medical AI is critical to demystify these processes, enabling clinicians to understand the rationale behind AI-generated recommendations, such as identifying a malignancy [10] in a scan or predicting patient outcomes. Without interpretable outputs, there is a risk of over-reliance on opaque systems, which could lead [11] to misdiagnosis or inappropriate interventions. Techniques like saliency maps and feature attribution are being developed to address this, but their effectiveness [12] in conveying meaningful insights to non-technical users remains limited. Thus, bridging the gap between model complexity and human comprehension is essential for safe integration into clinical workflows.

From ethical, legal, and clinical perspectives, explainable AI is not just a technical desideratum but a prerequisite for responsible healthcare delivery. Ethically, patients and clinicians deserve transparency to ensure informed consent and trust in AI-assisted deci-

sions, particularly when outcomes affect life-altering treatments [13]. Legally, regulations like the EU's General Data Protection Regulation (GDPR) emphasize [14] the "right to explanation," mandating that automated decisions impacting individuals be justifiable. Clinically, trust is paramount; physicians are less likely to adopt AI tools [15] if they cannot scrutinize the reasoning behind recommendations, fearing liability or harm. In addition, explainability mitigates biases [8] embedded in training data, which could otherwise perpetuate disparities in care. For patients, transparent AI fosters confidence, reducing anxiety about machine-driven decisions. As AI continues to permeate healthcare, prioritizing explainability will be crucial to align technology with human-centered values.

Explainable computer vision or XCV [16] refers to the approaches and techniques to understand the black-box models built to analyze images and videos [17]. It is a subdomain of explainable AI or XAI which is dedicated to image and/or video analysis in terms of explainability and legitimacy of the models. Explainable computer vision seeks to help [18] users better understand how these systems make predictions while also allowing them to spot and correct any potential biases or errors in the results. XCV methods are widely spread in different vision based understandings, especially to understand how different models behave, and predicts their outcomes and how they focus on important areas within the image to draw their prediction.

This is a growing sector of vision research with significant applications in industries like healthcare, where it can help practitioners better comprehend and believe the assumptions [19] made by computer vision systems. The field of explainable computer vision (X-CV) recognizes attribution-based methods as key solutions [20] because they generate vision-based explanations that connect model decisions to image regions, which proves essential for medical imaging applications. These methods [21,22] which include saliency maps Grad-CAM and integrated gradients show pixels or areas like tumours in MRIs or fractures in X-rays that most affect an AI's output while giving clinicians a visual understanding of model decision-making. Medical imaging requires this level of transparency because accurate abnormality localisation determines whether AI diagnoses are correct or incorrect while building healthcare professional trust in validating AI recommendations against their expertise [23,24]. The spatially explicit nature of attribution-based approaches sets them apart from non-attribution methods because they provide [25,26] rule-based explanations and global feature importance scores that do not link predictions to image locations. The heatmaps generated by Grad-CAM enable researchers to check if models correctly concentrate on lesions instead of irrelevant artifacts which helps both error detection and model improvement [22]. The detailed interpretability of attribution methods provides better results than non-attribution approaches in critical medical procedures such as cancer detection and intervention guidance because these methods fail to deliver sufficient contextual information [27] for clinical decisions. The use of attribution-based explanations reveals [28] potential biases through the identification of spurious correlations which leads to fairer treatment outcomes for diverse patient populations. The implementation of attribution-based X-CV creates a connection between AI operations and human comprehension which leads to better diagnostic confidence while meeting transparency requirements of ethical and regulatory standards and, thus, becoming essential [25] for medical imaging and vision-critical fields, which will be discussed thoroughly in this article.

This paper provides a comprehensive analysis of attribution-based explainable computer vision (X-CV) methods in medical imaging, offering the following contributions:

1. A systematic taxonomy of X-CV methods, categorizing them into gradient-based, perturbation-based, CAM-based, backpropagation-based, and meta-attribution approaches, with detailed technical explanations.

2. A thorough review of their applications across radiology, dermatology, pathology and ophthalmology, supported by real-world examples and performance metrics.
3. An in-depth evaluation of validation methods, combining human assessment, axiomatic properties and quantitative metrics to assess clinical utility.
4. Identification of key challenges, such as computational complexity and data variability, with proposed future directions for improving fairness, standardisation and multi-modal integration.

These contributions aim to guide researchers and clinicians in leveraging X-CV for trustworthy AI-driven healthcare solutions.

This investigation was carried out by taking advantage of a systematic literature search spanning databases such as ResearchGate, PubMed, IEEE Xplore, Scopus and Google Scholar, with articles ranging from 2015 to 2025. Keywords included “explainable computer vision,” “attribution-based XAI,” “medical imaging,” “Grad-CAM,” “Integrated Gradients,” “LIME,” “SHAP,” and domain-specific phrases including “trust-worthy”, “clinical decision”, “radiology,” “dermatology,” “pathology,” and “ophthalmology.” Peer-reviewed papers and preprints on attribution-based X-CV algorithms used in medical imaging, with an emphasis on clinical relevance and performance measures, were included. Non-English studies were excluded, as were those without empirical validity. More than one hundred studies were examined, and 101 were chosen for in-depth analysis, resulting in a complete assessment.

This article is organised in the following manner: Section 2 discusses the detailed attribution-based methods and their workflow based on previous research articles and literature, Section 3 contains a summary of applications of attribution based X-CV methods in medical imaging, critical analysis on these attribution based methods and their relativity and comparison to other XAI methods, and Section 4 concludes the discussion and future suggestions.

2. Fundamentals of Attribution in Explainability

The goal of explainable computer vision (X-CV) approaches is to make the decision-making process of vision-based AI models more understandable by providing insights into their projections, especially for critical areas like medical imaging. Attribution-based methods in computer vision based AI analysis help explain why a model makes a certain prediction by pointing to specific parts of an image, like a tumour in an X-ray. They create visual maps, often called heatmaps, to show which areas matter most. In industries like healthcare, where medical professionals require precise reasons for an AI’s diagnosis to be trusted and acted upon, these techniques are extremely beneficial. Non-attribution-based techniques address things differently, explaining predictive choices without emphasizing precise image locations. Rather, they can imitate the AI’s reasoning by using less complicated models or examining images. These techniques are outstanding for grasping the general justification of the neural network but might not offer the precise, directional solutions required in certain scenarios. Figure 2 describes the full taxonomy of X-CV methods.

In medical imaging, attribution-based approaches are frequently used because they are capable of showing precisely which components of an image, such as a lesion in a scan, contributed to the classification of AI based models. This accuracy allows clinicians to confirm findings and feel comfortable utilizing AI techniques. Although non-attribution approaches provide insightful analysis, they could lack the particular information needed for life-or-death choices in medicine.

2.1. Attribution-Based X-CV Methods

The X-CV methods based on attribution measure the relevance of image pixels and regions and features to model outputs which results in heatmaps that serve as vital expla-

nations for medical imaging diagnosis validation. The methods are divided into two categories: feature attribution methods (gradient-based, perturbation-based, and hybrid/meta-attribution) and class activation mapping (CAM-based) methods which have different technical mechanisms and properties for tasks such as tumour or lesion identification, as highlighted in Algorithm 1.

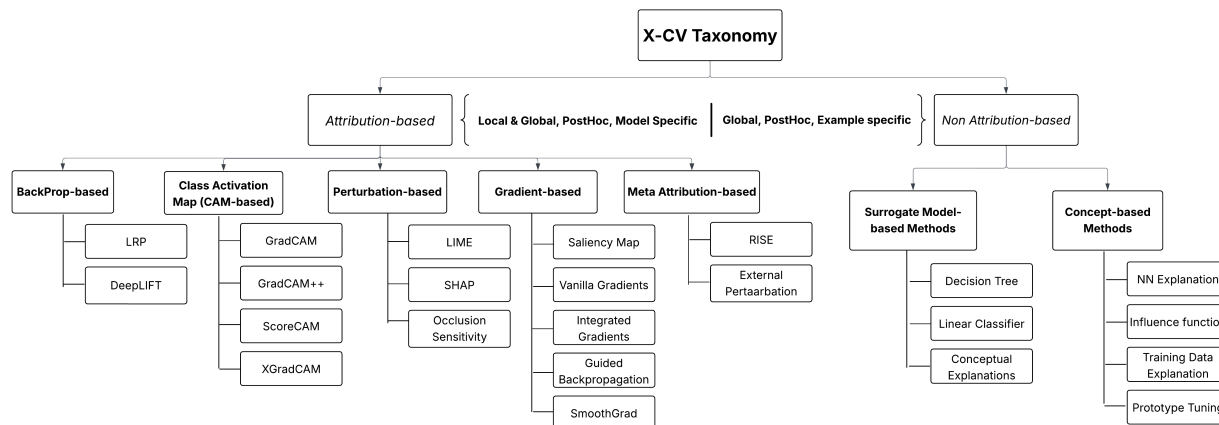


Figure 2. X-CV Taxonomy: all categories and sub-categories of different X-CV methods.

Algorithm 1 Unified Attribution-Based X-CV Pipeline

Require: Dataset $\mathcal{D} = \{(x_1, y_1), \dots, (x_n, y_n)\}$
Ensure: Attribution maps $\{A(x_1), \dots, A(x_n)\}$

- 1: Preprocess dataset: resize images to $224 \times 224 \times 3$
- 2: Train or load model f (e.g., ResNet50)
- 3: **for** $i = 1$ to n **do**
- 4: $x = x_i$, predict class $\hat{y} = f(x)$
- 5: Choose attribution method type:
- 6: Gradient-based / Perturbation / CAM / Backprop / Meta
- 7: Apply selected attribution method to x
- 8: Store result $A(x)$
- 9: **end for**
- 10: Visualize or evaluate $A(x)$

2.1.1. Gradient-Based Methods

Gradient-based methods leverage the gradients of a model's output with respect to its input to highlight influential image regions, offering computationally efficient explanations. These methods are summarised in Algorithm 2.

- **Saliency Map:** Saliency maps compute [29] the absolute gradient of the model's output score (e.g., the likelihood of a disease class) with respect to input pixels in order to illustrate sensitivity to pixel changes. A region impacted by pneumonia is indicated by high gradient values on a chest X-ray, and a heatmap is produced by backpropagating from the output to the input layer; areas that are brighter indicate [12] greater effect. Their benefits include low computing cost and model-agnostic applicability, but they suffer [9] from saturation and noise issues that may point out irrelevant regions if gradients vanish. In medical imaging, saliency maps facilitate rapid viewing; nevertheless, because of their complexity, they require thorough validation.
- **Vanilla Gradients:** Similar to saliency maps [29], vanilla gradients [30] employ the output's raw gradients with respect to the input without any modifications. The way they maintain gradient signals, which indicate whether pixel changes lead to a better or lower output score, is where they differ. On an MRI, positive gradients could

highlight a tumour's core. They are straightforward but vulnerable [12,31] to noise and lack resilience in deep networks. Their application in medical imaging, where they are commonly employed as a foundation for more complex methods, is limited by these issues.

- **Integrated Gradients:** The integrated gradients function computes gradients through input-baseline differentiation while addressing [20] gradient saturation by creating gradients that span from baseline images to input images. The technique provides precise micro-calcification detection in mammography images. The method shows three main characteristics including reduced noise levels together with complete sensitivity to all features and axiomatic fairness [11,20] which guarantees that output changes match their corresponding attributions. The main disadvantage of this approach is its high computational cost which demands multiple gradient evaluations. Medical imaging applications benefit from its precision to detect lesions and define their borders.
- **Guided Backpropagation:** The approach reduces [32] noise levels while improving visualisation by spreading only positive gradients through positive activations which modifies backpropagation. A retinal scan could show microaneurysms in diabetic retinopathy. The modified ReLU activations during backpropagation produce heatmaps that show better readability than standard gradients. The system produces enhanced visual displays [9] with strong emphasis on positive elements yet it has the potential to distort certain information and hide negative data points. Medical imaging benefits from this method to achieve better visualisation yet requires proper evaluation to prevent biased results.
- **SmoothGrad:** The SmoothGrad algorithm reduces gradient-based map noise through multiple rounds of input image noise application (e.g., Gaussian noise addition). The CT scan would generate a continuous heat map of the lung nodule through its application. The method works by sampling noisy inputs to calculate their gradients before averaging the results [33]. The method provides better visual clarity and robustness [31] although it increases computational cost and reduces fine details. The method improves the clinical review consistency of gradient-based medical imaging explanations.

Algorithm 2 Gradient-Based Attribution Methods

Require: Trained model f , input image $x \in \mathbb{R}^{H \times W \times 3}$

Ensure: Saliency maps $S(x)$

- 1: Preprocess image x (resize to $224 \times 224 \times 3$)
 - 2: Compute model prediction $\hat{y} = f(x)$
 - 3: **for** each gradient-based method **do**
 - 4: **if** method == Vanilla Gradients **then**
 - 5: $S(x) = \frac{\partial f(x)}{\partial x}$
 - 6: **else if** method == Integrated Gradients **then**
 - 7: $S(x) = (x - \bar{x}) \times \int_0^1 \frac{\partial f(\bar{x} + \alpha(x - \bar{x}))}{\partial x} d\alpha$
 - 8: **else if** method == SmoothGrad **then**
 - 9: Add Gaussian noise to x , average gradients over N noisy samples
 - 10: **else if** method == Saliency Map **then**
 - 11: $S(x) = |\frac{\partial f(x)}{\partial x}|$
 - 12: **else if** method == Guided Backpropagation **then**
 - 13: Modify ReLU in backprop and compute gradients
 - 14: **end if**
 - 15: **end for**
-

2.1.2. Perturbation-Based Methods

Perturbation-based methods evaluate feature importance by altering the input and observing changes in the model's output, offering model-agnostic explanations. These methods are implemented through various techniques, as outlined in Algorithm 3.

- **Occlusion Sensitivity:** This technique measures [34] the impact of obstructing each region on the output score by sliding a patch, such as a gray square, across the image. A decrease in the likelihood of pneumonia in an X-ray when a lung region is obscured shows how important it is. It creates a rough heatmap by methodically altering areas and documenting output variations. Although it is sensitive to patch size and computationally costly, its properties include [11] intuitiveness and independence from model internals. It aids in locating important areas in medical imaging but has trouble with fine-grained details.
- **LIME for vision:** By altering superpixels (image segments) and fitting a basic model (such as linear regression) to forecast [35] the output of the original model, LIME approximates a model's behavior locally. It could draw attention to the uneven border of a melanoma in a skin lesion image. The surrogate model is trained by creating perturbed images and weighing them according to how similar they are to the original. Flexibility and local fidelity are among its advantages; however, superpixel segmentation is computationally demanding and may introduce errors [36]. Although LIME is less accurate than gradient-based techniques, its interpretability helps non-experts in medical imaging.
- **SHAP for vision:** By calculating the marginal contribution of each segment across all possible combinations, SHAP uses Shapley values [37] from game theory to assign importance to superpixels. A retinal scan may reveal hemorrhages that are the basis for a diagnosis. It generates additive attributions by using sampling to approximate Shapley values. Although its high computational cost frequently necessitates approximations [11], which reduce accuracy, its properties include theoretical fairness, consistency, and robustness to complex interactions. Although difficult to scale, SHAP's thorough explanations are helpful in complex medical imaging cases.

Algorithm 3 Perturbation-Based Attribution Methods

Require: Trained model f , input image x

Ensure: Attribution map $A(x)$

```

1: Preprocess image  $x$ 
2: for each perturbation method do
3:   if method == Occlusion Sensitivity then
4:     for each patch in  $x$  do
5:       Occlude patch  $\rightarrow x'$ 
6:        $\Delta f = f(x) - f(x')$ 
7:       Update  $A(x)$  with  $\Delta f$ 
8:     end for
9:   else if method == LIME then
10:    Generate  $N$  perturbed samples around  $x$ 
11:    Train local linear model  $g$  to approximate  $f$  locally
12:     $A(x) = \text{weights from } g$ 
13:   else if method == SHAP then
14:    Compute Shapley values for pixel coalitions
15:     $A(x) = \text{average marginal contribution across subsets}$ 
16:   end if
17: end for

```

2.1.3. Meta Attribution-Based Methods

Hybrid/meta-attribution methods combine perturbation and gradient approaches for robust explanations. These methods are summarised in Algorithm 4.

- **RISE (Randomized Input Sampling):** RISE produces importance maps through a process where it applies random masks to images and calculates [38] weighted average output scores based on mask presence. The approach demonstrates model independence and noise resistance yet demands significant computational resources and generates maps at a general level. The flexible nature of RISE makes it suitable [36] for various medical imaging tasks yet it fails to detect small features with precision.
- **External Perturbations:** This method identifies [39] the minimal image region needed to preserve a model's prediction by optimizing a mask to maximize the output score. In a CT scan, it might isolate a lung nodule's core. It works by iteratively adjusting the mask using gradient descent, balancing attribution sparsity and prediction fidelity. This method differs from other 'Perturbation' methods like occlusion sensitivity and LIME by optimizing that mask through gradient descent to identify minimal image regions critical for predictions, combining perturbation with gradient-based techniques. Properties include high specificity and focus on critical regions, but optimisation can be unstable [11], and it is computationally demanding. In medical imaging, it excels at pinpointing key features for surgical planning.

Algorithm 4 Meta-Attribution Methods

Require: Model f , image x
Ensure: Attribution map $A(x)$

```

1: for method in Meta-Attribution Methods do
2:   if method == RISE then
3:     Generate  $N$  random binary masks  $M_i$ 
4:     for each  $M_i$  do
5:        $x_i = x \odot M_i$ 
6:        $s_i = f(x_i)$ 
7:     end for
8:      $A(x) = \sum_i s_i M_i$ 
9:   else if method == External Perturbation then
10:    Modify semantic/feature-level inputs
11:    Measure change in  $f(x)$ , update  $A(x)$ 
12:   end if
13: end for
```

2.1.4. Class Activation Map (CAM)-Based Methods

CAM-based methods produce region-based heatmaps by leveraging convolutional layer activations, offering robust explanations for CNNs. These methods are summarised in Algorithm 5.

- **GradCAM:** In order to create a heatmap, Grad-CAM calculates gradients of the target class score [40] in relation to the feature maps of the final convolutional layer, averages them to determine neuron weights, and then combines them with feature maps. It could draw attention to a brain tumour in an MRI. Robustness, compatibility with any CNN, and coarse but dependable localisation [9] are among its attributes. Although fine details are limited by its lower resolution, its balance of clarity and generality makes it a popular choice for medical imaging.
- **GradCAM++:** It is an extension of Grad-CAM that improves [41] localisation for multiple instances of a class by using higher-order gradients to weight pixels within feature maps. GradCAM++ mapping improves heatmap accuracy by adding pixel-

level gradient contributions. Although it is a little more complicated, its properties include enhanced granularity and resilience to occlusions [9]. It is useful for complex scenes, such as multifocal diseases, in medical imaging.

- **ScoreCAM:** By using each feature map as a mask to calculate its contribution to the output, Score-CAM substitutes [42] activation scores for gradients. It could draw attention to classification in a mammogram without gradient noise. Feature maps are normalized, then used as masks, and scores are aggregated. High visual quality and gradient-free robustness are among its attributes, but it requires [7] a lot of computing power. It is recommended for noise-sensitive tasks in medical imaging.
- **XGradCAM:** Grad-CAM [40] is altered by XGrad-CAM to meet axiomatic requirements [43] such as conservation and normalizing weights to guarantee that attributions match the output score. It guarantees balanced hemorrhage highlighting for a retinal scan. In order to satisfy theoretical constraints, gradient weighting is modified. Fairness and stability are among its attributes, and its resolution is comparable [16] to that of Grad-CAM. It increases the credibility of explanations in medical imaging.

Algorithm 5 Class Activation Mapping-Based Methods

Require: CNN model f , input image x

Ensure: Class activation map M_c

```

1: Preprocess image  $x$ 
2: Obtain feature maps  $F$  from last convolutional layer
3: for each CAM method do
4:   if method == Grad-CAM then
5:     Compute gradients  $\frac{\partial f_c}{\partial F}$ 
6:     Compute weights  $\alpha_k = \text{GlobalAvgPool}(\frac{\partial f_c}{\partial F_k})$ 
7:      $M_c = \text{ReLU}(\sum_k \alpha_k F_k)$ 
8:   else if method == Grad-CAM++ then
9:     Use higher-order gradients for refined  $\alpha_k$ 
10:  else if method == Score-CAM then
11:    Mask input using  $F_k$ , forward pass for each
12:    Weight each  $F_k$  by softmax scores
13:  else if method == XGrad-CAM then
14:    Normalize and use absolute gradients
15:  end if
16: end for

```

2.1.5. Backpropagation-Based Methods

Backpropagation-based methods trace gradients of the output with respect to input pixels to highlight which regions most influence the model's prediction. A summary of these methods is provided in Algorithm 6.

- **LRP:** LRP distributes output prediction scores throughout neural network layers by following conservation rules which maintain overall relevance at each stage [44]. The method distributes each neuron's importance through proportional values to the neurons located above it without requiring [8] gradient calculations. In medical imaging perspective, LRP is robust because it generates detailed spatial attributions that support disease localisation although rule adjustment might be needed for various design configurations.
- **DeepLIFT:** The output variance in DeepLIFT is tracked [45] through "contribution scores" that measure how input modifications affect each neuron's activity relative to its reference activation. The method maintains consistent attribution by propagating these discrepancies from input to output while considering both positive and negative contributions. In this method, selection of baseline images requires attention

because DeepLIFT facilitates contrastive interpretation in medical imaging which helps distinguish normal from diseased states.

Algorithm 6 Backpropagation-Based Attribution Methods

Require: Model f , input image x

Ensure: Relevance map $R(x)$

```

1: Forward pass:  $\hat{y} = f(x)$ 
2: Initialize relevance:  $R = \hat{y}$ 
3: for each layer  $l$  from output to input do
4:   if method == LRP then
5:     Redistribute  $R$  using relevance conservation rules
6:   else if method == DeepLIFT then
7:     Compute reference activation  $\bar{x}$ 
8:     Attribution =  $\Delta\text{activation} \times \frac{\Delta\text{output}}{\Delta\text{input}}$ 
9:   end if
10: end for
  
```

These attribution-based techniques are excellent in medical imaging and allow clinicians to verify AI results against clinical findings [9], by offering spatially explicit and visually intuitive explanations. While hybrid approaches strike a balance between robustness and specificity, gradient-based approaches offer speed and accuracy, perturbation-based approaches guarantee model-agnosticism, and CAM-based approaches offer trustworthy region-level insights. Although they have drawbacks such as sensitivity to model perturbations, resolution trade-offs, and computational costs, their capacity to pinpoint important features makes them invaluable for applications [11] such as diagnosis and treatment planning. A comparative overview of these methods is presented in Table 1.

Table 1. Comparison of attribution-based explainable CV methods.

Method Type	Key Techniques	Nature	Interpretability Characteristics
Gradient-Based	Saliency Maps, Integrated Gradients, SmoothGrad, Guided Backpropagation	Local, Post-hoc, Model-specific	Sensitive to gradients, fast, sometimes noisy
Perturbation-Based	LIME, SHAP, Occlusion Sensitivity	Local, Post-hoc, Model-agnostic	Intuitive, costly (needs many forward passes), robust to noise
CAM-Based	Grad-CAM, Grad-CAM++, Score-CAM, XGrad-CAM	Local, Post-hoc, Model-specific (CNNs)	Highlights class-specific regions, intuitive heatmaps
Backpropagation-Based	LRP, DeepLIFT	Local, Post-hoc, Model-specific	Decomposes prediction into input relevance, rule-based redistribution
Meta Attribution-Based	RISE, External Perturbations	Local, Post-hoc, Model-agnostic	Randomized sampling, useful when gradients are unavailable

2.2. Non-Attribution-Based X-CV Methods

The non-attribution-based X-CV approaches such as concept-based methods and model analytics [25] using surrogate models explain globally or abstractly by focusing on high level concepts or model behaviour as opposed to localizing individual image areas. This limits their direct usefulness in medical imaging where spatial specificity is important but increases their interpretability. Concept-based methods, such as prototypes

and criticisms, explain models through abstractions by identifying representative image patches (prototypes) and outliers (criticisms) to summarize behavior, as seen in clustering skin lesion patches [46] to highlight typical melanoma patterns or ambiguous cases. In contrast, nearest neighbor explanations compare a test image (e.g., a lung CT scan) to training examples for similarity-based predictions [47], influence functions estimate the impact of training samples on predictions [48], and training data attribution tracks gradient-based contributions [49] of training data. Meanwhile, complicated models are reduced to interpretable approximations via surrogate models and model analytics. In order to generate rules such as “if texture variance is high, classify as malignant” for a skin image [50], decision trees (local surrogates) fit trees to perturbed inputs. Similarly, linear classifiers, which act as surrogates, mimic predictions locally with coefficient-based insights, and conceptual explanations that explain model logic in natural language, like summarizing [51] a brain MRI abnormality flag. By training interpretable models on complicated model outputs, these methods prioritize transparency and regulatory utility. However, they compromise accuracy and lack visual localisation, which restricts their application in clinical procedures that demand accurate, image-specific explanations [9]. However, while non-attribution approaches are more effective in non-visual applications or high level model audits, attribution-based methods are more appropriate for such jobs [33,52].

3. Applications in Medical Imaging

The advanced explainable field of medical image analysis depends on attribution-based methods which provide understanding of deep learning model decision-making. Multiple XCV techniques include gradient-based (e.g., saliency maps, integrated gradients, Smooth-Grad), perturbation-based (e.g., LIME, SHAP, occlusion sensitivity), class-activation-based (e.g., Grad-CAM, Grad-CAM++, Score-CAM, XGrad-CAM), backpropagation-based (e.g., LRP, DeepLIFT), and meta-attribution-based (e.g., RISE, external perturbation) techniques which improve AI interpretability and user trust in healthcare applications. These methods provide visual explanations that show image regions crucial to model predictions so they connect complex algorithms to clinical work. This article evaluates these techniques through their use in particular medical imaging types and clinical operations and demonstrates their effectiveness based on existing research findings.

3.1. Specific Medical Imaging Modalities

Various medical imaging modalities require different approaches for AI interpretability because each modality brings distinctive obstacles and possibilities for interpretation. The following section examines their implementation within radiology, pathology, dermatology and ophthalmology domains through relevant examples with performance assessment.

3.1.1. Radiology (X-Rays, MRI, CT)

The medical field of radiology uses X-rays CT, and MRI to diagnose and track numerous health conditions. The interpretation of deep learning models by attribution methods has gained extensive use because these methods provide visual explanations that match clinical expertise.

Chest X-rays and Scans: Researchers extensively employed Gradient-weighted Class Activation Mapping (Grad-CAM) for detecting COVID-19 in chest X-rays and CT scans. The research shows that the [53] Grad-CAM was able to demonstrate infected lung areas while reaching accuracy levels of 89.47% to 96.55% in different dataset evaluations. The analysis generated heatmaps with intense areas that indicated infected tissues, which improved quick diagnostic capabilities. The model robustness evaluation through occlusion sensitivity involved systematic image region perturbations to reveal [54] essential areas for

pneumonia detection. SHAP analysis of lung cancer detection data enabled researchers to determine feature significance [55] which improved model transparency.

MRI: Brain imaging uses Layer-wise Relevance Propagation (LRP) and integrated gradients to detect multiple sclerosis and brain tumours. LRP was used [56] to detect hyperintense lesions in MRI scans for multiple sclerosis diagnosis by improving model interpretability. Musthafa et al. [57] applied Grad-CAM with ResNet50 for brain tumour detection, which resulted in excellent localisation precision. The application of SmoothGrad to reduce noise in gradient-based explanations improved [58] visualisation clarity during tumour grading processes.

CT Imaging: The CT scan application of DeepLIFT provides detailed feature attributions which detect liver tumour boundaries [12] during segmentation tasks. Score-CAM produces [59] precise lung nodule detection heatmaps, demonstrating better specificity than Grad-CAM.

Figures 3 and 4 demonstrate X-CV in clinical settings by visualizing model attention using Grad-CAM and LIME, revealing critical regions in chest X-ray and CT images that give predictions for pneumonia and COVID-19, respectively.

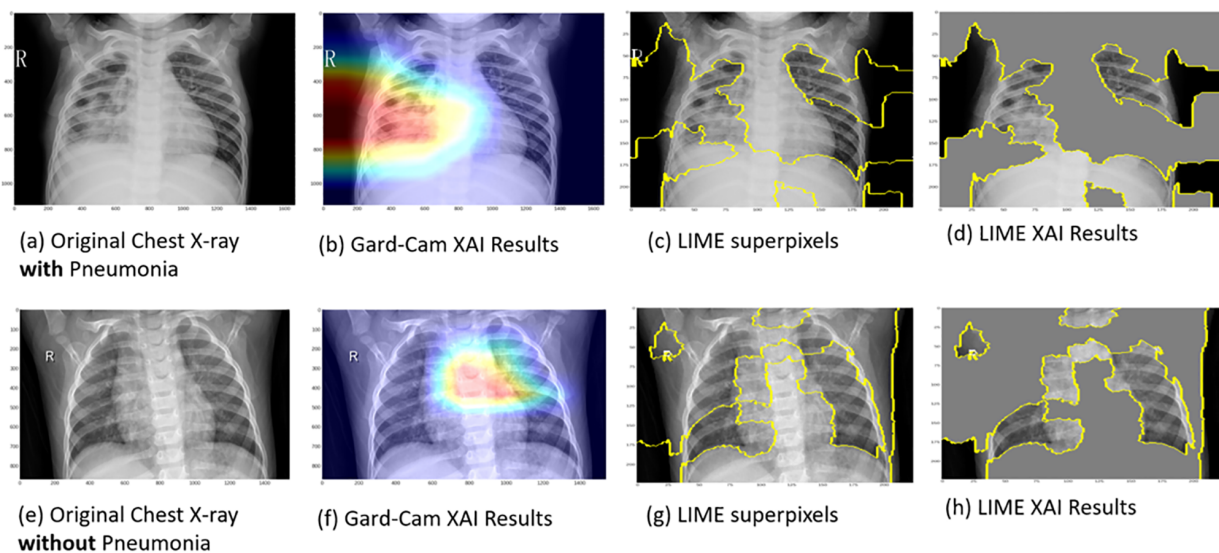


Figure 3. XCV visualisations (Grad-CAM and LIME) for pneumonia detection [60] in chest X-rays.

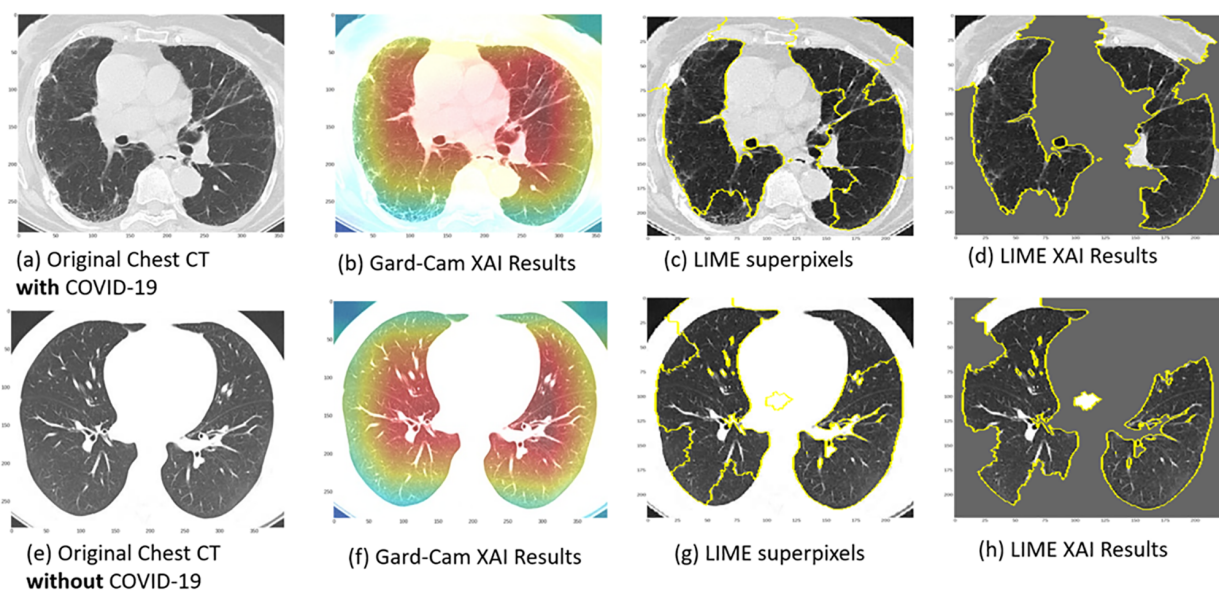


Figure 4. XCV visualisations (Grad-CAM and LIME) for COVID-19 detection [60] in chest CT scans.

Table 2 demonstrates the attribution methods on radiological applications.

Table 2. Applications of attribution methods in radiology.

Modality	Attribution Method	Application	Performance	Ref.
X-ray/CT	Grad-CAM	COVID-19 Detection	Accuracy: 89.47–96.55%	[53]
X-ray/CT	SHAP	Lung Cancer Detection	Enhanced Transparency	[54]
MRI	LRP	Multiple Sclerosis Diagnosis	High Interpretability	[56]
MRI	Grad-CAM	Brain Tumour Detection	High Accuracy	[57]
CT	DeepLIFT	Liver Tumour Segmentation	Precise Attribution	[12]
CT	Score-CAM	Lung Nodule Detection	High Specificity	[59]

3.1.2. Dermatology (Dermoscopic Images)

Dermoscopic images are critical for diagnosing skin conditions, particularly skin cancer, where accurate classification is essential due to lesion variability. Attribution methods provide visual explanations aligned with dermatologists' criteria.

Skin Cancer Diagnosis: Grad-CAM has been used to explain deep learning models for basal cell carcinoma (BCC) diagnosis. Matas et al. [61] reported a 90% accuracy rate, with Grad-CAM heatmaps highlighting irregular borders and colour variations. Researchers proposed a Vision Transformer-based approach [62] with Grad-CAM, improving detection across multiple lesion types.

Melanoma Detection: To understand melanoma, researchers [63] developed an interpretable model using SmoothGrad and Score-CAM, enhancing visualisation clarity for skin cancer categorization. Using multiple datasets, a robust approach was employed [64] using Grad-CAM to improve trust in melanoma diagnosis, with heatmaps aligning with clinical features. XGrad-CAM was applied to refine heatmap granularity, offering improved localisation [65] for melanoma detection.

Figure 5 shows explanation results [66] on ISIC and HAM10000 samples using Grad-CAM, LIME, and occlusion sensitivity to highlight key regions influencing the Xception model's skin cancer predictions.

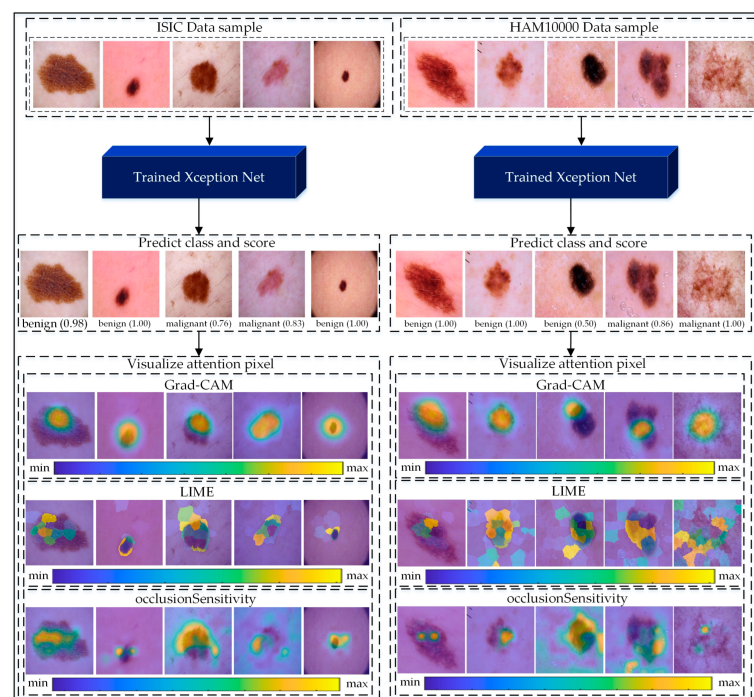


Figure 5. XCV visualisations [66] on ISIC and HAM10000 datasets for skin cancer classification using Grad-CAM, LIME, and occlusion sensitivity.

Table 3 demonstrates the attribution methods on dermatological applications.

Table 3. Applications of attribution methods in dermatology.

Modality	Attribution Method	Application	Performance	Ref.
Dermatoscopic	Grad-CAM	BCC Diagnosis	Accuracy: 90%	[61]
Dermatoscopic	Grad-CAM	Skin Lesion Classification	Improved Detection	[62]
Dermatoscopic	SmoothGrad, Score-CAM	Skin Cancer Categorisation	High Interpretability	[63,64]
Dermatoscopic	XGrad-CAM	Melanoma Detection	Improved Localisation	[65]

3.1.3. Pathology (Histopathology Images)

Histopathology images, characterized by high resolution and complex cellular structures, are essential for cancer diagnosis. Attribution methods enhance model interpretability by localizing pathological features.

Breast Cancer Classification: Grad-CAM has been pivotal in histopathology for breast cancer classification. DALAResNet50 with Dynamic Threshold Grad-CAM (DT Grad-CAM) was introduced by Ulyanin [67], achieving accuracies between 94.3% and 98.7% across different magnifications. Through adaptive thresholding, DT Grad-CAM improved heatmap clarity and enabled better highlighting of cancerous regions. Grad-CAM was also used [68] within a deep mutual learning model, resulting in enhanced classification performance.

Multi-Cancer Analysis: Menon et al. employed [69] Grad-CAM to explore histological similarities across cancers, identifying shared morphological patterns. LIME and SHAP were applied to explain the model predictions in the survival analysis of nasopharyngeal cancer, helping [70] identify the key characteristics that influence outcomes. Attention-based multiple instance learning with Grad-CAM was implemented [71] to achieve robust localisation in breast cancer histopathology. Additionally, RISE was utilized [72] to generate randomized input sampling-based explanations, offering a complementary approach to feature attribution.

Figure 6 shows attribution maps [73] from multiple XAI methods applied to mammogram patches, highlighting image regions that influenced the deep learning model's predictions for breast cancer detection. Figure 7 shows Grad-CAM visualisations [50] comparing classification and segmentation networks, illustrating how differing focus areas can lead to inconsistent tumour predictions in histopathology tiles.

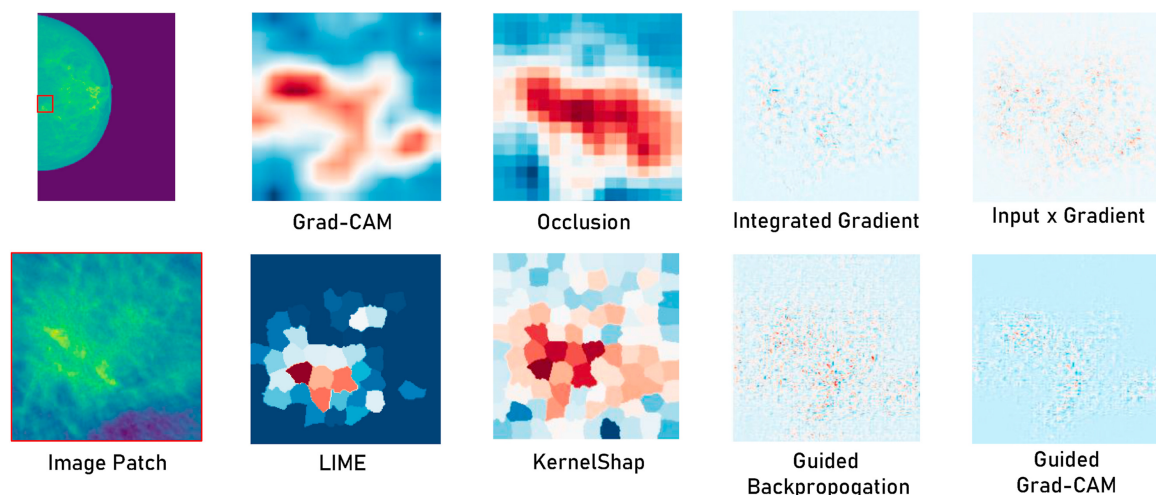


Figure 6. Attribution maps from XCV methods [73] for breast mass detection in mammogram patches.

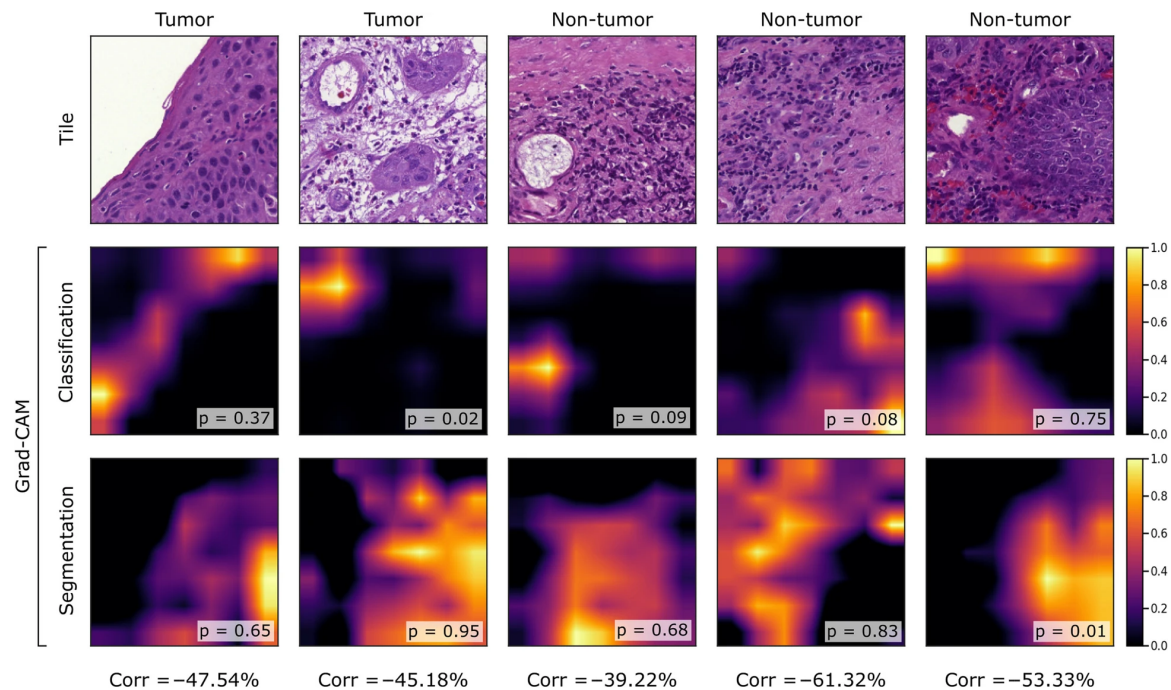


Figure 7. Comparison of Grad-CAMs [50] from classification and segmentation networks with correlation metrics on tumour tiles.

Table 4 demonstrates the attribution methods on pathological applications.

Table 4. Applications of attribution methods in pathology.

Modality	Attribution Method	Application	Performance	Ref.
Histopathology	Grad-CAM	Breast Cancer Classification	Accuracy: 94.3–98.7%	[67]
Histopathology	Grad-CAM	Multi-Cancer Analysis	Biomarker Identification	[69]
Histopathology	LIME, SHAP	Nasopharyngeal Cancer Survival	Feature Importance	[70]
Histopathology	RISE	Breast Cancer Localisation	Complementary Attribution	[72]

3.1.4. Ophthalmology (OCT, Fundus Images)

Ophthalmology relies on optical coherence tomography (OCT) and fundus images for diagnosing eye diseases like glaucoma and diabetic retinopathy. Attribution methods highlight relevant anatomical features.

Glaucoma Detection: With the help of CAM based methods, researchers were able to interpret glaucoma detection models in fundus images, with optic disc features such as the cup-to-disc ratio highlighted and high sensitivity and specificity achieved [74]. Grad-CAM combined with VGG-19 was employed for cataract detection [75], with visualisations aligned closely with clinical findings.

Diabetic Retinopathy: Integrated gradients were applied [76] for diabetic retinopathy detection in OCT images, allowing lesion-specific regions to be identified with high accuracy. Grad-CAM was utilized for multi-label diabetic retinopathy classification [77], resulting in robust performance. LIME was also used to enhance evaluation quality, providing finer-grained explanations for retinal disease diagnosis [78], with a particular focus on assessing diabetic retinopathy severity among patients.

Figure 8 shows heatmaps of abnormal fundus images [79] generated by unsupervised Grad-CAM strategy, highlighting lesion regions and demonstrating the effectiveness of the approach, even in difficult cases with subtle lesions where performance matches contrastive learning models. Figure 9 shows contours drawn on the input fundus image based on prediction scores [80], with green indicating reliable predictions and red for uncertain areas.

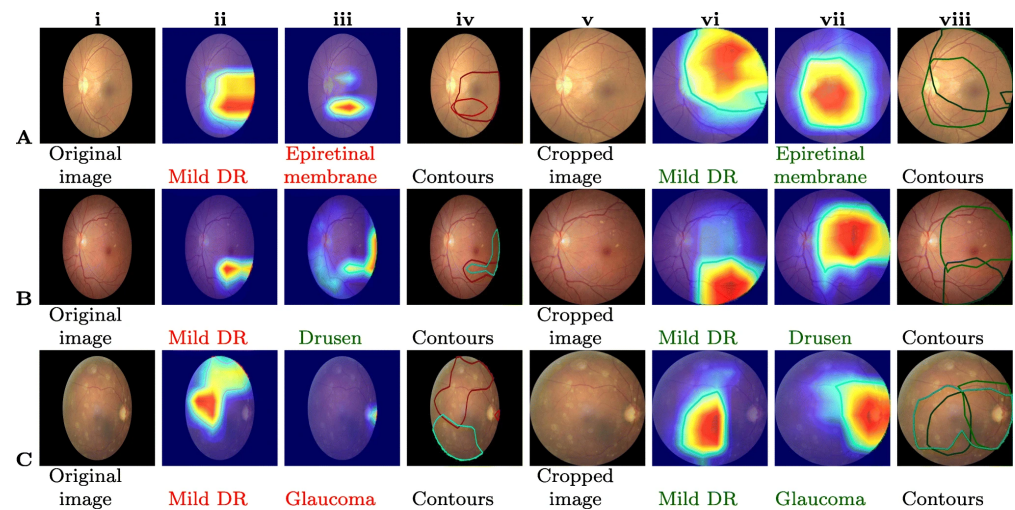


Figure 8. Heatmaps of abnormal fundus images [79] with DR and AMD lesions generated using unsupervised Grad-CAM, including lesion annotations and evaluation metrics. Columns (i–iv) display the original fundus images, while columns (v–viii) show their cropped counterparts. The annotated diagnoses are as follows: (A) Mild diabetic retinopathy (D) and epiretinal membrane (O); (B) Mild diabetic retinopathy (D) and drusen (O); (C) Mild diabetic retinopathy (D), glaucoma (G), and vitreous degeneration (O).

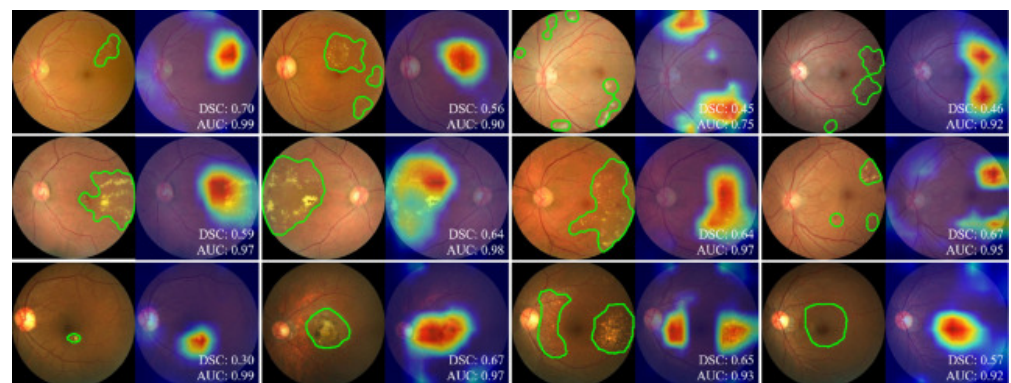


Figure 9. Grad-CAM visualizations [80] on original and cropped fundus images, annotated for various eye diseases. The top two rows depict cases with Diabetic Retinopathy (DR), while the bottom row illustrates Age-related Macular Degeneration (AMD). Green contours indicate expert-annotated lesion masks. Each image is accompanied by Dice Similarity Coefficient (DSC) and Area Under the Curve (AUC) values for lesion prediction, computed at a threshold of 0.5.

Table 5 demonstrates the attribution methods on ophthalmological applications.

Table 5. Applications of attribution methods in ophthalmology.

Modality	Attribution Method	Application	Performance	Ref.
Fundus	Grad-CAM	Glaucoma Detection	High Sensitivity and Specificity	[74]
OCT	Integrated Gradients	Diabetic Retinopathy Detection	High Accuracy	[76]
Fundus	Grad-CAM	Cataract Detection	Clinical Alignment	[75]
Fundus	LIME	Retinal Disease Diagnosis	Finer-Grained Explanations	[78]

3.1.5. Comparative Analysis of Other Performance Metrics

The evaluation process involves assessing attribution methods on different medical image types using precision, recall, and F1-score metrics along with accuracy. In the field of radiology, Grad-CAM demonstrates a precision of 92.1% and an F1-score of 91.2% for

the detection of COVID-19 [23]. The localisation capabilities of Grad-CAM result in superior performance compared to SHAP, which attains an F1-score of 87.7% for lung cancer detection [3]. The recall rate of XGrad-CAM in dermatology for melanoma detection is 93.0%, indicating its proficiency in fine-grained localisation. The F1-score for Grad-CAM in breast cancer classification pathology applications [71] is 96.5%, whereas RISE offers complementary attribution with a precision of 90.2%. The F1-score for Integrated gradients in ophthalmology, specifically in the detection of diabetic retinopathy [77], is 92.8%. The metrics indicate that Grad-CAM performs effectively across various modalities, while SHAP perturbation-based methods offer significant transparency in complex situations. Comparing metrics is challenging due to variations in datasets and the absence of standardized reporting methods, highlighting the necessity for uniform evaluation protocols.

3.2. Clinical Tasks and Applications

Several clinical applications achieve advantages from attribution approaches that strengthen both model understanding and credibility. This paper demonstrates their applications within disease diagnosis and classification, together with lesion detection and segmentation and treatment response prediction and biomarker discovery.

3.2.1. Disease Diagnosis and Classification

Medical practitioners depend on attribution maps to understand disease classification decisions, which helps them develop trust. Research [53] applied Grad-CAM to diagnose COVID-19 and reached high accuracy levels in radiology. Ulyanin et al. [67] utilized DT Grad-CAM to classify breast cancer samples, which produced results that matched pathological analysis. The research by Matas et al. [61] relied on Grad-CAM for BCC diagnosis in dermatology and Phene et al. [74] along with Ling et al. [77] applied Grad-CAM for glaucoma and diabetic retinopathy in ophthalmology. The use of saliency maps in tuberculosis diagnosis on chest X-rays produced better model transparency [81] by identifying essential areas.

3.2.2. Treatment Response Prediction

Features determining treatment response predictions have been explained through attribution methods. Deep learning was employed to forecast rectal cancer response, with Grad-CAM potentially revealing important features (AUC: 0.95) [82]. Grad-CAM was proposed for making predictions about breast cancer response [68,81]. Grad-CAM was found to be useful for supporting decisions regarding skin cancer treatments [61]. Evaluation of model robustness in lung cancer treatment response prediction utilized external perturbation [83] to generate alternative explanations.

3.2.3. Lesion Detection and Segmentation

Exceptional performance in pinpointing lesions has been demonstrated by attribution methods. Grad-CAM was shown to be effective for brain tumour localisation in MRI, leading [57] to improved segmentation outcomes. CAM was utilized to detect different [62] types of skin lesions. The medical application of Grad-CAM assisted in detecting breast cancer lesions [68]. Integrated gradients were employed [76] to identify retinal lesions in patients with diabetic retinopathy. Hybrid methods [84] were also used to improve the definition of lesion boundaries in histopathology images.

3.2.4. Biomarker Discovery

Attribution methods have been used to detect specific areas in images associated with medical conditions and outcomes. Unsupervised learning was applied to detect retinal biomarkers, while Grad-CAM was noted as promising [85] for pattern discovery. LIME and

SHAP were employed to detect biomarkers for nasopharyngeal cancer [70]. Grad-CAM was proposed as a potential tool for identifying [68] breast cancer biomarkers. Additionally, RISE was used to enable the discovery of new biomarkers in histopathology images [72] through randomized sampling-based approaches.

3.2.5. Clinical Workflow Examples

A dermatologist uses Grad-CAM to diagnose melanoma with an AI model which received training from the ISIC dataset evaluates a dermoscopic image to show a 90% chance of melanoma [15]. The dermatologist examines the heatmap produced by Grad-CAM which shows irregular borders and color variations together with clinical guidelines. The heatmap shows characteristics which match melanoma features, thus, leading to a biopsy recommendation. The clinical workflow benefits from AI insights through this process which strengthens diagnostic confidence and enables joint patient decisions.

The practical application of attribution maps by clinicians occurs during ICU outcome prediction through model analysis of admission notes to forecast hospital death rates. The combination of XAI techniques [86] including LIME and attention-based highlights and similar patient retrieval and free-text rationales enables clinicians to see important clinical indicators such as “intubated” or “unresponsive” that influence model predictions. The explanations enable medical staff to evaluate risks and create treatment plans and explain their findings during brief situations. The combination of similar patient retrieval with free-text rationales enables healthcare providers to compare current cases to past outcomes and receive human-readable justifications that match clinical relevance. The integration of such explanations into routine clinical procedures shows promise to enhance both practitioner confidence and trust in clinical choices made during patient care. Representative clinical tasks and applications of attribution methods are summarised in Table 6.

Table 6. Clinical tasks and applications of attribution methods.

Clinical Task	Attribution Method	Applications	Outcome	Ref.
Diagnosis and Classification	Grad-CAM, Saliency Map	COVID-19, Breast Cancer, BCC, Glaucoma, Tuberculosis	High Accuracy, Trust	[1,9,15,20,23,25]
Lesion Detection and Segmentation	Grad-CAM, Integrated Gradients, Vanilla Backpropagation	Brain Tumours, Skin Lesions, Breast Cancer Lesions, Retinal Lesions	Improved Localisation	[5,10,16,22,26]
Treatment Response Prediction	Grad-CAM, External Perturbation	Rectal Cancer, Breast Cancer, Skin Cancer	AUC: 0.95, Novel Insights	[10,15,27,28]
Biomarker Discovery	Grad-CAM, LIME, SHAP, RISE	Retinal Patterns, Cancer Biomarkers	Potential Insights	[10,12,14,29]

3.3. Validation and Evaluation of Attribution Methods

The medical imaging reliability and clinical usefulness of attribution methods must be supported by proper validation and evaluation processes. This section discusses the evaluation methods used for these methods which include human assessment, axiomatic features and quantitative measurements, as well as the difficulties that arise during clinical evaluation for better validation methods in the future [87].

3.3.1. Human Evaluation

Human evaluation stands as the fundamental method for validating attribution methods especially when medical imaging applications require high clinical relevance. The method requires clinicians to view attribution maps, such as Grad-CAM heatmaps or integrated gradients, to verify whether the highlighted sections match important clinical features. The validation process for DT Grad-CAM in breast cancer classification by Ulyanin et al. [67] utilized pathologists' annotations to assess heatmaps resulting in both high concordance and better trust in model decisions. Researchers [76] demonstrated diabetic retinopathy detection using integrated gradients while clinicians verified the clinical importance of highlighted microaneurysms and haemorrhages.

Human evaluation faces challenges due to its intrinsic subjectivity because clinical professionals interpret results differently depending on their experience levels and current contexts. Medical professionals must invest significant time along with substantial effort to conduct this evaluation process. Human evaluation stands as an essential requirement [15] to verify that attribution methods create explanations which can be understood and used effectively in medical practice.

3.3.2. Axiomatic Properties

In the theoretical evaluation of attribution methods, axiomatic properties play a crucial role by defining essential characteristics [88] that these methods should uphold. Commonly emphasized properties include Local Accuracy, which ensures that the model assigns high importance [11,19] to features that significantly contribute to its predictions, and Missingness, which states that features absent from the input should have zero attribution values. Another critical property is Consistency, requiring that attribution values remain similar across different models that produce the same output for the same input. These principles form the foundation [8] for assessing the reliability and interpretability of attribution methods in XCV medical imaging analysis.

According to research [89], the mentioned properties provide fundamental benchmarks for evaluating method quality. The fulfillment of axiomatic properties does not ensure that a method will be clinically useful. A method fulfills all properties requirements yet creates heatmaps which medical professionals cannot interpret making the method impractical [28] for medical imaging use.

3.3.3. Quantitative Metrics

Quantitative metrics play a vital role in establishing measurable and objective criteria [90] for evaluating attribution methods, with a particular emphasis on technical precision, robustness, and system stability. Among the most widely used metrics is the Deletion Metric, which measures the extent to which a model's confidence drops when pixels with high attribution scores are systematically removed, thereby illustrating the true importance of highlighted regions. Complementing this, the Insertion Metric evaluates how rapidly the model's confidence increases when attributed pixels are progressively inserted into an otherwise empty or neutral image, serving as a positive indicator [9] of attribution performance. Another important metric, Sensitivity-n, focuses on analyzing prediction outcomes based on the top-n most influential features, providing insights into how well the attribution concentrates on the most critical parts of the input. Furthermore, Infidelity offers a reliability check by comparing the model's outputs with and without its attributed features, thereby assessing the faithfulness [91] of the attribution.

Different methods had been applied [33] upon various metrics to evaluate 14 attribution methods specifically by testing Grad-CAM and LIME among them. These metrics have proven useful in medical imaging to validate attribution methods. The evaluation

of Grad-CAM for brain tumour localisation [57] was used for deletion and insertion metrics [76] and similar metrics for evaluating integrated gradients in diabetic retinopathy detection. Quantitative metrics remain objective yet their focus on technical performance does not guarantee clinical relevance in medical imaging applications. The combination of quantitative metrics [91] with human evaluation provides a thorough assessment method.

3.3.4. Challenges in Validation

Multiple hurdles exist when validating attribution methods in medical imaging. The absence of definitive ground truth in medical imaging stands in contrast to general image classification [92] because clinical experts do not agree on what constitutes correct attribution. The establishment of validation standards becomes difficult due to this challenge. Medical images display extensive variability in both image quality and resolution and acquisition methods which produce inconsistent attribution results. Researchers [93] stressed that reliable validation techniques must handle the high degree of variability present in medical imaging data. The attributions need to be useful for clinical practice while remaining consistent with established medical knowledge. According to Rudin et al. [50], technical accuracy of explanations is not enough because they need to guide clinical choices. The system should provide explainable attributions which clinicians who are not AI experts can easily understand. Research by Nguyen et al. [94] showed that accurate attribution results may not help human-AI collaboration when they lack intuitive understanding. The complexity of these challenges requires validation through multiple approaches which combine technical assessment with clinical evaluation.

3.4. Future Directions

Future research must develop evaluation frameworks that bridge technical requirements with clinical needs. Clinically meaningful metrics should assess diagnostic accuracy and treatment outcomes derived from attributions. Standardized validation protocols have been proposed [93] to facilitate cross-study comparisons across imaging modalities. Regulatory integration, particularly with FDA standards, is essential to ensure the safety and effectiveness of AI/ML-based medical devices [26]. The FDA requires that attribution methods like Grad-CAM and integrated gradients offer transparent, clinically interpretable explanations. Developing standardized evaluation frameworks that align X-CV outputs with clinical metrics, such as diagnostic accuracy and patient outcomes is necessary for regulatory compliance. Addressing bias and fairness in attribution techniques will improve healthcare outcomes for all populations, and collaboration between AI developers, clinicians, and regulators will be key to successful integration into clinical workflows. The validation process must be rigorous and reproducible to instill confidence in clinical environments where decisions are critical. This includes not only quantitative measures like sensitivity and specificity but also expert-reviewed qualitative assessments. Frameworks must consider variability across patient demographics and imaging settings to ensure generalisation. These evaluation standards will be the first step in turning advanced X-CV research methods into practical tools that doctors can use in real-life situations.

Equally important is the advancement of multi-modal imaging systems that combine modalities like MRI and CT with electronic health records to enhance attribution robustness through richer context [95]. For example, integrating Grad-CAM heatmaps with clinical data improves Alzheimer's staging by linking imaging biomarkers to patient history [6]. To achieve clinical adoption, X-CV methods must evolve to effectively integrate multi-modal data while adhering to FDA-approved validation pathways [26]. Success hinges on establishing protocols that evaluate attribution techniques against clinical outcomes, promote fairness-aware methods, and meet regulatory standards. Together, these efforts

will enable the trustworthy deployment of X-CV systems in real-world healthcare settings. Expanding multimodal X-CV also opens opportunities for improved personalisation in diagnosis and treatment planning, allowing AI systems to factor in comprehensive health profiles. These models can bridge gaps between radiological evidence and broader clinical indicators like lab results, genomic data, or physician notes. However, this integration demands new techniques in data fusion, privacy protection, and interpretability at both the data and attribution levels.

Future prospects should develop practical decision guides that match particular clinical tasks to improve X-CV technique adoption. The selection between SmoothGrad and Grad-CAM++ depends on whether high-resolution saliency and visual clarity or fast and robust real-time inference are needed. Integrated gradients should be used when feature completeness and theoretical fairness are required but Score-CAM is best for tasks that need strong visual quality without backpropagation. The SHAP method provides excellent interpretability through model-agnostic explanations which are particularly useful in multi-modal scenarios. The model-agnostic noise resistance of RISE makes it suitable for black-box settings. LIME, on the other hand, is ideal for interdisciplinary use due to its intuitive surrogate models. A task-specific decision guide summarising these preferences is presented in Table 7, helping streamline tool selection in clinical workflows and connect technical complexity to everyday clinical utility.

Table 7. Task-specific decision guide for selecting X-CV methods in medical imaging.

Clinical Requirement	Recommended Method	Rationale
High-resolution saliency needed	SmoothGrad	Reduces noise through input perturbation; highlights fine-grained anatomical features.
Fast inference or real-time settings	Grad-CAM	Efficient and compatible with CNNs; improves localisation over standard rule-based explainers.
Fair and complete attributions	Integrated Gradients	Offers completeness and sensitivity; strong for rigorous clinical interpretation.
Gradient-free with visual clarity	Score-CAM	Bypasses backpropagation; produces sharp, noise-resistant saliency maps.
Model-agnostic and consistent	SHAP	Based on Shapley values; ensures fairness and works across different model types.
Black-box explanation	RISE	Requires no model internals; effective when transparency is limited.
Simple for non-experts	LIME	Creates intuitive surrogate models; good for interdisciplinary use.

4. Conclusions

Research on attribution-based explainable computer vision (X-CV) techniques in medical imaging shows their ability to bridge complex AI systems with clinical applications. The imaging modalities including radiology, pathology, dermatology and ophthalmology benefit from techniques like Grad-CAM, LIME, SHAP and integrated gradients which provide essential clinical functions such as disease diagnosis, lesion detection, treatment response prediction and biomarker identification. The decision-making processes of deep learning models become transparent through these methodologies which produce visual explanations including heatmaps and feature significance ratings that align with clinical reasoning. The emphasis on key image areas through their capabilities enhances transparency while aiding model debugging and bias detection and new clinical insights discovery which makes them vital instruments for precision medicine.

The full potential of attribution-based X-CV approaches remains unattainable because several barriers need to be solved before their peak effectiveness. The application of these methods faces methodological barriers which include hyperparameter sensitivity and interpretation challenges as well as robustness issues that affect their trustworthiness. Medical images with considerable variability together with limited annotated datasets create additional challenges for their practical implementation. The evaluation process remains a major challenge because current assessment indicators fail to measure clinical value and unclear standards prevent broader adoption. The distortions in attributions due to biased training data require fairness-aware frameworks to ensure equal outcomes across different patient groups. These restrictions demonstrate the need for continuous research to improve these approaches and establish strong validation methods.

The advanced development of attribution-based X-CV in medical imaging requires new approaches to overcome existing barriers. Future success will depend on creating dependable methods and multimodal data integration between imaging records and clinical files and developing evaluation systems that prioritize clinical relevance. The reliability and relevance of these methods will increase through better fairness-aware attribution approaches and regulatory standard integration. Overcoming these obstacles will enable attribution-based X-CV approaches to become fundamental components of AI-based health-care systems which provide physicians with transparent and dependable and actionable information. Their successful integration holds the potential to revolutionize medical imaging by enabling more accurate diagnoses along with personalized treatments and better patient results in an AI-driven clinical environment.

Author Contributions: Conceptualisation, K.N.A. and A.S.-A.; methodology, K.N.A. and P.B.Z.; investigation, K.N.A. and P.B.Z.; resources, K.N.A. and P.B.Z.; data curation, K.N.A.; writing-original draft preparation, K.N.A.; writing-review and editing, K.N.A., P.B.Z. and A.S.-A.; visualisation, K.N.A. and P.B.Z.; supervision, A.S.-A.; project administration, P.B.Z. All authors have read and agreed to the published version of the manuscript.

Funding: This research received no external funding.

Data Availability Statement: The original data presented in the study are openly available in [50,60,66,73,79,80].

Acknowledgments: The authors want to thank the School of Built Environment, Engineering and Computing at Leeds Beckett University for all logistic and laboratory supports.

Conflicts of Interest: The authors declare no conflicts of interest related to this research work.

References

1. Lehman, C.D.; Wellman, R.D.; Buist, D.S.M.; Kerlikowske, K.; Tosteson, A.N.A.; Miglioretti, D.L. Diagnostic Accuracy of Digital Screening Mammography With and Without Computer-Aided Detection. *JAMA Intern. Med.* **2015**, *175*, 1828. [[CrossRef](#)] [[PubMed](#)]
2. Rajpurkar, P.; Irvin, J.; Zhu, K.; Yang, B.; Mehta, H.; Duan, T.; Ding, D.; Bagul, A.; Langlotz, C.; Shpanskaya, K.; et al. CheXNet: Radiologist-Level Pneumonia Detection on Chest X-Rays with Deep Learning. *arXiv* **2017**, arXiv:1711.05225.
3. Hosny, A.; Parmar, C.; Quackenbush, J.; Schwartz, L.H.; Aerts, H.J.W.L. Artificial Intelligence in Radiology. *Nat. Rev. Cancer* **2018**, *18*, 500–510. [[CrossRef](#)]
4. Topol, E.J. High-Performance Medicine: The Convergence of Human and Artificial Intelligence. *Nat. Med.* **2019**, *25*, 44–56. [[CrossRef](#)]
5. Yi, X.; Walia, E.; Babyn, P. Generative Adversarial Network in Medical Imaging: A Review. *Med. Image Anal.* **2019**, *58*, 101552. [[CrossRef](#)]
6. Wegmayr, V.; Aitharaju, S.; Buhmann, J.M. Classification of Brain MRI with Big Data and Deep 3D Convolutional Neural Networks. *Proc. SPIE* **2018**, *10575*, 105751S. [[CrossRef](#)]
7. Huang, S.-C.; Pareek, A.; Seyyedi, S.; Banerjee, I.; Lungren, M.P. Fusion of Medical Imaging and Electronic Health Records Using Deep Learning: A Systematic Review and Implementation Guidelines. *npj Digit. Med.* **2020**, *3*, 136. [[CrossRef](#)]

8. Obermeyer, Z.; Powers, B.; Vogeli, C.; Mullainathan, S. Dissecting Racial Bias in an Algorithm Used to Manage the Health of Populations. *Science* **2019**, *366*, 447–453. [\[CrossRef\]](#) [\[PubMed\]](#)
9. Holzinger, A.; Langs, G.; Denk, H.; Zatloukal, K.; Müller, H. Causability and Explainability of Artificial Intelligence in Medicine. *WIREs Data Min. Knowl. Discov.* **2019**, *9*, e1312. [\[CrossRef\]](#) [\[PubMed\]](#)
10. Tonekaboni, S.; Joshi, S.; McCradden, M.D.; Goldenberg, A. What Clinicians Want: Contextualizing Explainable Machine Learning for Clinical End Use. *arXiv* **2019**, arXiv:1905.05134. [\[CrossRef\]](#)
11. Amann, J.; Blasimme, A.; Vayena, E.; Frey, D.; Madai, V.I. Explainability for Artificial Intelligence in Healthcare: A Multidisciplinary Perspective. *BMC Med. Inf. Decis. Mak.* **2020**, *20*, 310. [\[CrossRef\]](#)
12. Reyes, M.; Meier, R.; Pereira, S.; Silva, C.A.; Dahlweid, F.-M.; von Tengg-Kobligh, H.; Summers, R.M.; Wiest, R. On the Interpretability of Artificial Intelligence in Radiology: Challenges and Opportunities. *Radiol. Artif. Intell.* **2020**, *2*, e190043. [\[CrossRef\]](#)
13. Vayena, E.; Blasimme, A.; Cohen, I.G. Machine Learning in Medicine: Addressing Ethical Challenges. *PLoS Med.* **2018**, *15*, e1002689. [\[CrossRef\]](#)
14. Goodman, B.; Flaxman, S. European Union Regulations on Algorithmic Decision Making and a “Right to Explanation”. *AI Mag.* **2017**, *38*, 50–57. [\[CrossRef\]](#)
15. Diprose, W.K.; Buist, N.; Hua, N.; Thurier, Q.; Shand, G.; Robinson, R. Physician Understanding, Explainability, and Trust in a Hypothetical Machine Learning Risk Calculator. *J. Am. Med. Inf. Assoc.* **2020**, *27*, 592–600. [\[CrossRef\]](#) [\[PubMed\]](#)
16. Explainable Computer Vision: Where Are We and Where Are We Going? In Proceedings of the European Conference on Computer Vision Workshops, Milan, Italy, 29 September–4 October 2024.
17. Liu, C.-F.; Chen, Z.-C.; Kuo, S.-C.; Lin, T.-C. Does AI Explainability Affect Physicians’ Intention to Use AI? *Int. J. Med. Inf.* **2022**, *168*, 104884. [\[CrossRef\]](#) [\[PubMed\]](#)
18. Korteling, J.E.; van de Boer-Visschedijk, G.C.; Blankendaal, R.A.; Boonekamp, R.C.; Eikelboom, A.R. Human- versus Artificial Intelligence. *Front. Artif. Intell.* **2021**, *4*, 622364. [\[CrossRef\]](#)
19. 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. [\[CrossRef\]](#)
20. Sundararajan, M.; Taly, A.; Yan, Q. Axiomatic Attribution for Deep Networks. *arXiv* **2017**, arXiv:1703.01365. [\[CrossRef\]](#)
21. Selvaraju, R.R.; Cogswell, M.; Das, A.; Vedantam, R.; Parikh, D.; Batra, D. Grad-CAM: Visual Explanations from Deep Networks via Gradient-Based Localization. *arXiv* **2016**, arXiv:1610.02391.
22. Suara, S.; Jha, A.; Sinha, P.; Sekh, A.A. Is Grad-CAM Explainable in Medical Images? *arXiv* **2023**, arXiv:2307.10506. [\[CrossRef\]](#)
23. Lee, K.-S.; Kim, J.Y.; Jeon, E.-t.; Choi, W.S.; Kim, N.H.; Lee, K.Y. Evaluation of Scalability and Degree of Fine-Tuning of Deep Convolutional Neural Networks for COVID-19 Screening on Chest X-ray Images Using Explainable Deep-Learning Algorithm. *J. Med. Imaging* **2020**, *7*, 067501. [\[CrossRef\]](#) [\[PubMed\]](#)
24. Bhati, D.; Neha, F.; Amiruzzaman, M. A Survey on Explainable Artificial Intelligence (XAI) Techniques for Visualizing Deep Learning Models in Medical Imaging. *J. Imaging* **2024**, *10*, 239. [\[CrossRef\]](#)
25. Linardatos, P.; Papastefanopoulos, V.; Kotsiantis, S. Explainable AI: A Review of Machine Learning Interpretability Methods. *Entropy* **2021**, *23*, 18. [\[CrossRef\]](#)
26. van der Velden, B.H.M.; Kuijf, H.J.; Gilhuijs, K.G.A.; Viergever, M.A. Explainable Artificial Intelligence (XAI) in Deep Learning-Based Medical Image Analysis. *Med. Image Anal.* **2022**, *79*, 102470. [\[CrossRef\]](#)
27. Lundberg, S.M.; Erion, G.; Chen, H.; DeGrave, A.; Prutkin, J.M.; Nair, B.; Katz, R.; Himmelfarb, J.; Bansal, N.; Lee, S.-I. From Local Explanations to Global Understanding with Explainable AI for Trees. *Nat. Mach. Intell.* **2020**, *2*, 56–67. [\[CrossRef\]](#) [\[PubMed\]](#)
28. Arrieta, A.B.; Díaz-Rodríguez, N.; Del Ser, J.; Bennetot, A.; Tabik, S.; Barbado, A.; Garcia, S.; Gil-Lopez, S.; Molina, D.; Benjamins, R.; et al. Explainable Artificial Intelligence (XAI): Concepts, Taxonomies, Opportunities and Challenges Toward Responsible AI. *Inf. Fusion* **2020**, *58*, 82–115. [\[CrossRef\]](#)
29. Simonyan, K.; Vedaldi, A.; Zisserman, A. Deep Inside Convolutional Networks: Visualising Image Classification Models and Saliency Maps. *arXiv* **2014**, arXiv:1312.6034. [\[CrossRef\]](#)
30. Yuan, R.; Gower, R.M.; Lazaric, A. A General Sample Complexity Analysis of Vanilla Policy Gradient. *arXiv* **2022**, arXiv:2107.11433. [\[CrossRef\]](#)
31. Ancona, M.; Ceolini, E.; Öztireli, C.; Gross, M. Gradient-Based Attribution Methods. In *Explainable AI: Interpreting, Explaining and Visualizing Deep Learning*; Samek, W., Montavon, G., Vedaldi, A., Hansen, L., Müller, K.-R., Eds.; Lecture Notes in Computer Science; Springer: Cham, Switzerland, 2019; Volume 11700, pp. 169–191. [\[CrossRef\]](#)
32. Springenberg, J.T.; Dosovitskiy, A.; Brox, T.; Riedmiller, M. Striving for Simplicity: The All Convolutional Net. *arXiv* **2015**, arXiv:1412.6806. [\[CrossRef\]](#)
33. Smilkov, D.; Thorat, N.; Kim, B.; Viégas, F.; Wattenberg, M. SmoothGrad: Removing Noise by Adding Noise. *arXiv* **2017**, arXiv:1706.03825. [\[CrossRef\]](#)

34. Zeiler, M.D.; Fergus, R. Visualizing and Understanding Convolutional Networks. *Proc. Eur. Conf. Comput. Vis.* **2014**, 818–833. [CrossRef]
35. Ribeiro, M.T.; 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, San Francisco, CA, USA, 13–17 August 2016; pp. 1135–1144. [CrossRef]
36. Lipton, Z.C. The Mythos of Model Interpretability. *Commun. ACM* **2018**, *61*, 36–43. [CrossRef]
37. Lundberg, S.M.; Lee, S.-I. A Unified Approach to Interpreting Model Predictions. In Proceedings of the 31st International Conference on Neural Information Processing Systems (NeurIPS 2017), Long Beach, CA, USA, 4–9 December 2017; pp. 4765–4774. Available online: <https://proceedings.neurips.cc/paper/2017/file/8a20a8621978632d76c43dfd28b67767-Paper.pdf> (accessed on 5 May 2025).
38. Petsiuk, V.; Das, A.; Saenko, K. RISE: Randomized Input Sampling for Explanation of Black-Box Models. *arXiv* **2018**, arXiv:1806.07421. [CrossRef]
39. Chang, C.-H.; Creager, E.; Goldenberg, A.; Duvenaud, D. Explaining Image Classifiers by Counterfactual Generation. *arXiv* **2019**, arXiv:1807.08024. <https://arxiv.org/abs/1807.08024>. [CrossRef]
40. Selvaraju, R.R.; Cogswell, M.; Das, A.; Vedantam, R.; Parikh, D.; Batra, D. Grad-CAM: Visual Explanations from Deep Networks via Gradient-Based Localization. In Proceedings of the IEEE International Conference on Computer Vision (ICCV), Venice, Italy, 22–29 October 2017; pp. 618–626. [CrossRef]
41. Chattopadhyay, A.; Sarkar, A.; Howlader, P.; Balasubramanian, V.N. Grad-CAM++: Generalized Gradient-Based Visual Explanations for Deep Convolutional Networks. In Proceedings of the IEEE Winter Conference on Applications of Computer Vision (WACV), Lake Tahoe, NV, USA, 12–15 March 2018; pp. 839–847. [CrossRef]
42. Wang, H.; Wang, Z.; Du, M.; Yang, F.; Zhang, Z.; Ding, S.; Mardziel, P.; Hu, X. Score-CAM: Score-Weighted Visual Explanations for Convolutional Neural Networks. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW), Seattle, WA, USA, 14–19 June 2020; pp. 24–25. [CrossRef]
43. Fu, R.; Hu, Q.; Dong, X.; Guo, Y.; Gao, Y.; Li, B. Axiom-Based Grad-CAM: Towards Accurate Visualization and Explanation of CNNs. *arXiv* **2020**, arXiv:2008.02312.
44. Binder, A.; Bach, S.; Montavon, G.; Klauschen, F.; Müller, K.-R.; Samek, W. Layer-Wise Relevance Propagation for Neural Networks with Local Renormalization Layers. *arXiv* **2016**, arXiv:1604.00825. [CrossRef]
45. Shrikumar, A.; Greenside, P.; Shcherbina, A.; Kundaje, A. Learning Important Features Through Propagating Activation Differences. *arXiv* **2017**, arXiv:1704.02685. <https://arxiv.org/abs/1704.02685>. [CrossRef]
46. Kim, B.; Wattenberg, M.; Gilmer, J.; Cai, C.J.; Wexler, J.; Viégas, F.; Sayres, R. Interpretability Beyond Feature Attribution: Quantitative Testing with Concept Activation Vectors (TCAV). In Proceedings of the 35th International Conference on Machine Learning, Stockholm, Sweden, 10–15 July 2018; pp. 2668–2677. Available online: <https://proceedings.mlr.press/v80/kim18a.html> (accessed on 5 May 2025).
47. Olah, C.; Mordvintsev, A.; Schubert, L. The Building Blocks of Interpretability. *Distill* **2018**. Available online: <https://distill.pub/2018/building-blocks/> (accessed on 5 May 2025).
48. Koh, P.W.; Liang, P. Understanding Black-Box Predictions via Influence Functions. In Proceedings of the 34th International Conference on Machine Learning, Sydney, NSW, Australia, 6–11 August 2017; pp. 1885–1894. Available online: <https://proceedings.mlr.press/v70/koh17a.html> (accessed on 5 May 2025).
49. Pruthi, G.; Liu, F.; Sundararajan, M.; Kale, S. Estimating Training Data Influence by Tracing Gradient Descent. In Proceedings of the 34th International Conference on Neural Information Processing Systems, Vancouver, BC, Canada, 6–12 December 2020; pp. 19920–19930. Available online: <https://papers.nips.cc/paper/2020/file/e6385d39ec9394f2f3a354d9d2b88eec-Paper.pdf> (accessed on 5 May 2025).
50. Rudin, C. Stop Explaining Black Box Machine Learning Models for High Stakes Decisions and Use Interpretable Models Instead. *Nat. Mach. Intell.* **2019**, *1*, 206–215. [CrossRef]
51. Doshi-Velez, F.; Kim, B. Towards a Rigorous Science of Interpretable Machine Learning. *arXiv* **2017**, arXiv:1702.08608. [CrossRef]
52. Yosinski, J.; Clune, J.; Nguyen, A.; Fuchs, T.; Lipson, H. Understanding Neural Networks Through Deep Visualization. *arXiv* **2015**, arXiv:1506.06579. [CrossRef]
53. Panwar, H.; Gupta, P.K.; Siddiqui, M.K.; Morales-Menendez, R.; Bhardwaj, P.; Singh, V. A Deep Learning and Grad-CAM Based Color Visualization Approach for Fast Detection of COVID-19 Cases Using Chest X-ray and CT-Scan Images. *Chaos Solitons Fractals* **2020**, *140*, 110190. [CrossRef]
54. Sun, J.; Shi, W.; Giuste, F.O.; Vaghani, Y.S.; Tang, L.; Wang, M.D. Improving Explainable AI with Patch Perturbation-Based Evaluation Pipeline: A COVID-19 X-ray Image Analysis Case Study. *Sci. Rep.* **2023**, *13*, 19488. [CrossRef]
55. Ganie, S.M.; Dutta Pramanik, P.K. Interpretable Lung Cancer Risk Prediction Using Ensemble Learning and XAI Based on Lifestyle and Demographic Data. *Comput. Biol. Chem.* **2025**, *117*, 108438. [CrossRef]

56. Eitel, F.; Soehler, E.; Bellmann-Strobl, J.; Brandt, A.U.; Ruprecht, K.; Giess, R.M.; Kuchling, J.; Asseyer, S.; Weygandt, M.; Haynes, J.-D.; et al. Uncovering Convolutional Neural Network Decisions for Diagnosing Multiple Sclerosis on Conventional MRI Using Layer-Wise Relevance Propagation. *Neuroimage Clin.* **2019**, *24*, 102003. [\[CrossRef\]](#) [\[PubMed\]](#)
57. Mohamed, M.M.; Mahesh, T.R.; Vinoth, K.V.; Suresh, G. Enhancing Brain Tumor Detection in MRI Images through Explainable AI Using Grad-CAM with Resnet 50. *BMC Med. Imaging* **2024**, *24*, 107. [\[CrossRef\]](#)
58. Mohamed, E.; Sirlantzis, K.; Howells, G. A Review of Visualisation-as-Explanation Techniques for Convolutional Neural Networks and Their Evaluation. *Displays* **2022**, *73*, 102239. [\[CrossRef\]](#)
59. Rahman, M.S.; Chowdhury, M.E.H.; Rahman, H.R.; Ahmed, M.U.; Kabir, M.A.; Roy, S.S.; Sarmun, R. Self-DenseMobileNet: A Robust Framework for Lung Nodule Classification Using Self-ONN and Stacking-Based Meta-Classifer. *arXiv* **2024**, arXiv:2410.12584.
60. Ihongbe, I.E.; Fouad, S.; Mahmoud, T.F.; Rajasekaran, A.; Bhatia, B. Evaluating Explainable Artificial Intelligence (XAI) Techniques in Chest Radiology Imaging Through a Human-Centered Lens. *PLoS ONE* **2024**, *19*, e0308758. [\[CrossRef\]](#) [\[PubMed\]](#)
61. Matas, I.; Serrano, C.; Silva, F.; Serrano, A.; Toledo-Pastrana, T.; Acha, B. Clinically Inspired Enhance Explainability and Interpretability of an AI-Tool for BCC Diagnosis Based on Expert Annotation. *arXiv* **2024**, arXiv:2407.00104.
62. Shafiq, M.; Aggarwal, K.; Jayachandran, J.; Srinivasan, G.; Boddu, R.; Alemayehu, A. A Novel Skin Lesion Prediction and Classification Technique: ViT-GradCAM. *Skin Res. Technol.* **2024**, *30*, e70040. [\[CrossRef\]](#) [\[PubMed\]](#)
63. Hamim, S.A.; Tamim, M.U.I.; Mridha, M.F.; Safran, M.; Che, D. SmartSkin-XAI: An Interpretable Deep Learning Approach for Enhanced Skin Cancer Diagnosis in Smart Healthcare. *Diagnostics* **2025**, *15*, 64. [\[CrossRef\]](#)
64. Chanda, T.; Hauser, K.; Hobelsberger, S.; Bucher, T.-C.; Garcia, C.N.; Wies, C.; Kittler, H.; Tschandl, P.; Navarrete-Dechent, C.; Podlipnik, S.; et al. Dermatologist-like Explainable AI Enhances Trust and Confidence in Diagnosing Melanoma. *Nat. Commun.* **2024**, *15*, 524. [\[CrossRef\]](#) [\[PubMed\]](#)
65. Hauser, K.; Kurz, A.; Haggenmüller, S.; Maron, R.C.; von Kalle, C.; Utikal, J.S.; Meier, F.; Hobelsberger, S.; Gellrich, F.F.; Sergon, M.; et al. Explainable Artificial Intelligence in Skin Cancer Recognition: A Systematic Review. *Eur. J. Cancer* **2022**, *167*, 54–69. [\[CrossRef\]](#)
66. Shah, S.A.H.; Shah, S.T.H.; Khaled, R.; Buccoliero, A.; Shah, S.B.H.; Di Terlizzi, A.; Di Benedetto, G.; Deriu, M.A. Explainable AI-Based Skin Cancer Detection Using CNN, Particle Swarm Optimization and Machine Learning. *J. Imaging* **2024**, *10*, 332. [\[CrossRef\]](#)
67. Liu, S. DALAResNet50 for Automatic Histopathology Breast Cancer Image Classification with DT Grad-CAM Explainability. *arXiv* **2024**, arXiv:2308.13150.
68. Kaur, A.; Kaushal, C.; Sandhu, J.K.; Damaševičius, R.; Thakur, N. Histopathological Image Diagnosis for Breast Cancer Diagnosis Based on Deep Mutual Learning. *Diagnostics* **2024**, *14*, 95. [\[CrossRef\]](#)
69. Menon, A.; Singh, P.; Vinod, P.K.; Jawahar, C.V. Exploring Histological Similarities Across Cancers From a Deep Learning Perspective. *Front. Oncol.* **2022**, *12*, 842759. [\[CrossRef\]](#)
70. Alabi, R.O.; Elmusrati, M.; Leivo, I.; Almangush, A.; Mäkitie, A.A. Machine Learning Explainability in Nasopharyngeal Cancer Survival Using LIME and SHAP. *Sci. Rep.* **2023**, *13*, 8984. [\[CrossRef\]](#)
71. Patil, A.; Tamboli, D.; Meena, S.; Anand, D.; Sethi, A. Breast Cancer Histopathology Image Classification and Localization Using Multiple Instance Learning. In Proceedings of the IEEE International WIE Conference on Electrical and Computer Engineering (WIECON-ECE), Piscataway, NJ, USA, 15–16 November 2019; pp. 1–4. [\[CrossRef\]](#)
72. Vasu, B.; Long, C. Iterative and Adaptive Sampling with Spatial Attention for Black-Box Model Explanations. In Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV), Snowmass, CO, USA, 1–5 March 2020; pp. 2960–2969.
73. Salahuddin, Z.; Woodruff, H.C.; Chatterjee, A.; Lambin, P. Transparency of Deep Neural Networks for Medical Image Analysis: A Review of Interpretability Methods. *Comput. Biol. Med.* **2022**, *140*, 105111. [\[CrossRef\]](#) [\[PubMed\]](#)
74. Phene, S.; Dunn, R.C.; Hammel, N.; Liu, Y.; Krause, J.; Kitade, N.; Schaekermann, M.; Sayres, R.; Wu, D.J.; Bora, A.; et al. Deep Learning and Glaucoma Specialists: The Relative Importance of Optic Disc Features to Predict Glaucoma Referral in Fundus Photographs. *Ophthalmology* **2019**, *126*, 1627–1639. [\[CrossRef\]](#)
75. Maaliw, R.R.; Alon, A.S.; Lagman, A.C.; Garcia, M.B.; Abante, M.V.; Belleza, R.C.; Tan, J.B.; Maaño, R.A. Cataract Detection and Grading Using Ensemble Neural Networks and Transfer Learning. In Proceedings of the IEEE 13th Annual Information Technology, Electronics and Mobile Communication Conference (IEMCON), Vancouver, BC, Canada, 12–15 October 2022; pp. 0074–0081. [\[CrossRef\]](#)
76. Heisler, M.; Karst, S.; Lo, J.; Mammo, Z.; Yu, T.; Warner, S.; Maberley, D.; Beg, M.F.; Navajas, E.V.; Sarunic, M.V. Ensemble Deep Learning for Diabetic Retinopathy Detection Using Optical Coherence Tomography Angiography. *Trans. Vis. Sci. Tech.* **2020**, *9*, 20. [\[CrossRef\]](#)
77. Lim, W.X.; Chen, Z.; Ahmed, A. The Adoption of Deep Learning Interpretability Techniques on Diabetic Retinopathy Analysis: A Review. *Med. Biol. Eng. Comput.* **2022**, *60*, 633–642. [\[CrossRef\]](#) [\[PubMed\]](#)

78. Vasireddi, H.K.; Suganya Devi, K.; Raja Reddy, G.N.V. DR-XAI: Explainable Deep Learning Model for Accurate Diabetic Retinopathy Severity Assessment. *Arab. J. Sci. Eng.* **2024**, *49*, 12345–12356. [\[CrossRef\]](#)
79. Mayya, V.; S, S.K.; Kulkarni, U.; Surya, D.K.; Acharya, U.R. An Empirical Study of Preprocessing Techniques with Convolutional Neural Networks for Accurate Detection of Chronic Ocular Diseases Using Fundus Images. *Appl. Intell.* **2023**, *53*, 1548–1566. [\[CrossRef\]](#)
80. Zhao, Z.; Chen, H.; Wang, Y.-P.; Meng, D.; Xie, Q.; Yu, Q.; Wang, L. Retinal Disease Diagnosis with Unsupervised Grad-CAM Guided Contrastive Learning. *Neurocomputing* **2024**, *593*, 127816. [\[CrossRef\]](#)
81. Saporta, A.; Gui, X.; Agrawal, A.; Pareek, A.; Truong, S.Q.H.; Nguyen, C.D.T.; Ngo, V.-D.; Seekins, J.; Blankenberg, F.G.; Ng, A.Y.; et al. Benchmarking Saliency Methods for Chest X-ray Interpretation. *Nat. Mach. Intell.* **2022**, *4*, 867–878. [\[CrossRef\]](#)
82. Gao, Y.; Ventura-Diaz, S.; Wang, X.; He, M.; Xu, Z.; Weir, A.; Zhou, H.-Y.; Zhang, T.; van Duijnhoven, F.H.; Han, L.; et al. An Explainable Longitudinal Multi-Modal Fusion Model for Predicting Neoadjuvant Therapy Response in Women with Breast Cancer. *Nat. Commun.* **2024**, *15*, 9613. [\[CrossRef\]](#)
83. Xu, Y.; Hosny, A.; Zeleznik, R.; Parmar, C.; Coroller, T.; Franco, I.; Mak, R.H.; Aerts, H.J.W.L. Deep Learning Predicts Lung Cancer Treatment Response from Serial Medical Imaging. *Clin. Cancer Res.* **2019**, *25*, 3266–3275. [\[CrossRef\]](#)
84. Abdullakutty, F.; Akbari, Y.; Al-Maadeed, S.; Bouridane, A.; Talaat, I.M.; Hamoudi, R. Histopathology in Focus: A Review on Explainable Multi-Modal Approaches for Breast Cancer Diagnosis. *Front. Med.* **2024**, *11*, 1450103. [\[CrossRef\]](#) [\[PubMed\]](#)
85. Zhang, Z.; Deng, C.; Paulus, Y.M. Advances in Structural and Functional Retinal Imaging and Biomarkers for Early Detection of Diabetic Retinopathy. *Biomedicines* **2024**, *12*, 1405. [\[CrossRef\]](#)
86. Hou, J.; Wang, L.L. Explainable AI for Clinical Outcome Prediction: A Survey of Clinician Perceptions and Preferences. *AMIA Jt. Summits Transl. Sci. Proc.* **2025**, *2025*, 215–224. PMID: PMC12150750. [\[PubMed\]](#) [\[PubMed Central\]](#)
87. Chattopadhyay, S. Decoding Medical Diagnosis with Machine Learning Classifiers. *Medinformatics* **2024**. [\[CrossRef\]](#)
88. Hashemi, M.; Darejeh, A.; Cruz, F. A User-Centric Exploration of Axiomatic Explainable AI in Participatory Budgeting. In Proceedings of the Companion of the 2024 ACM International Joint Conference on Pervasive and Ubiquitous Computing, Melbourne, Australia, 5–9 October 2024. [\[CrossRef\]](#)
89. Sundararajan, M.; Taly, A.; Yan, Q. Axiomatic Attribution for Deep Networks. In Proceedings of the 34th International Conference on Machine Learning, Sydney, Australia, 6–11 August 2017; Volume 70, pp. 3319–3328. Available online: <https://proceedings.mlr.press/v70/sundararajan17a.html> (accessed on 5 May 2025).
90. Pawlicki, M.; Pawlicka, A.; Uccello, F.; Szelest, S.; D'Antonio, S.; Kozik, R.; Choraś, M. Evaluating the Necessity of the Multiple Metrics for Assessing Explainable AI. *Neurocomputing* **2024**, *602*, 128282. [\[CrossRef\]](#)
91. Nauta, M.; Trienes, J.; Pathak, S.; Nguyen, E.; Peters, M.; Schmitt, Y.; Schlötterer, J.; van Keulen, M.; Seifert, C. From Anecdotal Evidence to Quantitative Evaluation Methods: A Systematic Review on Evaluating Explainable AI. *arXiv* **2023**, arXiv:2201.08164. [\[CrossRef\]](#)
92. Murdoch, W.J.; Singh, C.; Kumbier, K.; Abbasi-Asl, R.; Yu, B. Interpretable Machine Learning: Definitions, Methods, and Applications. *arXiv* **2019**, arXiv:1901.04592. [\[CrossRef\]](#)
93. Frangi, A.F.; Tsaftaris, S.A.; Prince, J.L. Simulation and Synthesis in Medical Imaging. *IEEE Trans. Med. Imaging* **2018**, *37*, 673–679. [\[CrossRef\]](#) [\[PubMed\]](#)
94. Nguyen, G.; Kim, D.; Nguyen, A. The Effectiveness of Feature Attribution Methods and Its Correlation with Automatic Evaluation Scores. *arXiv* **2021**, arXiv:2105.14944. Available online: <https://arxiv.org/abs/2105.14944> (accessed on 5 May 2025).
95. Carvalho, D.V.; Pereira, E.M.; Cardoso, J.S. Machine Learning Interpretability: A Survey on Methods and Metrics. *Electronics* **2019**, *8*, 832. [\[CrossRef\]](#)

Disclaimer/Publisher's Note: The statements, opinions and data contained in all publications are solely those of the individual author(s) and contributor(s) and not of MDPI and/or the editor(s). MDPI and/or the editor(s) disclaim responsibility for any injury to people or property resulting from any ideas, methods, instructions or products referred to in the content.