

Review

Explainable Artificial Intelligence for Skin Lesion Classification: A Comprehensive Review of Methods and Challenges

Jennifer Whewell , Rebecca Peters  and Janusz Kulon * 

Faculty of Computing, Engineering and Science, University of South Wales, Pontypridd CF37 1DL, UK; jennifer.whewell@southwales.ac.uk (J.W.); rebecca.peters@southwales.ac.uk (R.P.)

* Correspondence: j.kulon@southwales.ac.uk

Abstract

The rapid advancement of machine learning and artificial intelligence (AI) has created new opportunities to enhance diagnostic accuracy in dermatology, particularly within primary care settings. Computer-aided diagnosis (CAD) systems have demonstrated potential to support General Practitioners (GPs) by enabling earlier and more consistent identification of skin diseases. This review critically examines the literature on explainable artificial intelligence (XAI) for skin disease classification, with a specific focus on the evolution of explainability frameworks and the methodological implications of dataset selection. A comprehensive review of studies published between 2020 and 2025 was conducted across multiple academic databases, encompassing research on skin lesion detection, classification, and monitoring. The analysis reveals that deep learning architectures, particularly those leveraging transfer learning with models such as EfficientNet, ResNet, and Xception, frequently report high classification accuracies—often exceeding 90% when evaluated on single benchmark datasets. However, studies employing multiple datasets consistently demonstrate more stable and generalisable performance, albeit with modest reductions in reported accuracy, highlighting a critical trade-off between performance optimisation and real-world robustness. The review further identifies a clear temporal progression in the adoption of XAI techniques. Early studies relied on a broader range of post hoc explainability while later work increasingly consolidated around Grad-CAM, SHAP, and related attribution techniques, followed by gradual diversification into more specialised frameworks such as TCAVs (Testing with Concept Activation Vectors) and Prototype-based Networks. Despite these advances, the lack of clinically grounded explanations, limited integration of ethical considerations, and reliance on non-clinical imagery continue to constrain clinical applicability which we have explored using a GRADE-style narrative. Notably, evidence suggests that CAD systems can improve GP diagnostic accuracy for conditions such as melanoma and seborrhoeic keratosis; however, sustained clinical adoption remains contingent on transparent, reliable, and context-aware explainability mechanisms.



Academic Editor: Hongying Meng

Received: 24 March 2026

Revised: 15 June 2026

Accepted: 18 June 2026

Published: 25 June 2026

Copyright: © 2026 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 \(CC BY\)](https://creativecommons.org/licenses/by/4.0/) license.

Keywords: XAI; explainability; deep learning; interpretability

1. Introduction

Skin conditions are among the most common reasons for consultation in primary care, resulting in over one million dermatology referrals annually in the UK [1]. Approximately 50% of these referrals relate to suspected skin malignancies [2]; however, only 6% are ultimately confirmed as malignant [3]. This discrepancy highlights the diagnostic challenges faced by general practitioners (GPs), whose limited dermatological training can contribute

to diagnostic uncertainty and unnecessary referrals [4]. Harkemanne et al. [5] found that GPs without specific melanoma training demonstrated inadequate diagnostic accuracy in several high-risk clinical scenarios. Consequently, improving the accuracy of skin lesion assessment remains an important clinical priority.

Understanding skin lesions is critical for improving diagnostic precision. Skin lesions can be classified into two categories: primary and secondary. Primary lesions are atypical dermatological changes that may be congenital or develop over time, while secondary lesions result from changes in primary lesions due to external factors. Both types can be broadly defined as abnormal growths or appearances of the skin in relation to the surrounding skin [6]. The current assessment of skin lesions in primary care commonly relies on clinical criteria such as the ABCD rule (Asymmetry, Border, Colour, Diameter). However, lesion presentations vary considerably, and some melanomas do not exhibit typical ABCD characteristics, making accurate diagnosis difficult [7]. Distinguishing malignant lesions from benign conditions such as seborrheic keratosis further complicates referral decisions and may contribute to both over-referral and delayed diagnosis.

Advances in AI have created new opportunities for CAD in dermatology. Studies suggest that AI systems can support clinical decision-making in primary care and, in some cases, outperform less experienced clinicians in melanoma detection [7,8]. Furthermore, AI-assisted approaches have shown potential to reduce diagnostic errors and improve clinical decision-making [9,10]. Despite these promising results, most studies to date evaluate models using retrospective image datasets and offline validation, with comparatively limited exploration of prospective clinical deployment or real-time decision support in routine dermatological practice. A small number of prospective and clinician-in-the-loop studies have been conducted. For example, previous work has explored case studies for clinician-facing explainable AI approaches [11], while Yu et al. compared performance across clinical and dermoscopic images to better understand modality-specific diagnostic accuracy [12]. Nevertheless, such studies remain limited in scale and are not yet representative of widespread clinical integration. Broader smart healthcare frameworks have also been proposed to facilitate AI adoption in clinical settings [13].

While high diagnostic accuracy is essential, successful clinical adoption requires AI systems to be transparent and interpretable. Many high-performing machine learning models, particularly deep learning architectures, operate as “black boxes”, making it difficult for clinicians to understand how predictions are generated [14]. This lack of transparency poses significant challenges for clinicians, who must not only trust but also understand the reasoning behind a model’s predictions in order to incorporate them into diagnostic and treatment decisions. A central issue in this regard is the interpretability–performance paradox: the observed trade-off between model complexity and the ability to understand model decision-making processes [15]. In clinical applications, this paradox necessitates a careful balance between predictive performance and interpretability to ensure that AI systems are not only accurate but also transparent and usable in practice. Consequently, there is a growing recognition of the need for transparent and interpretable AI systems that provide clear justifications for their outputs, thereby fostering trust, accountability, and reliability within clinical practice.

This challenge has led to growing interest in Explainable Artificial Intelligence (XAI), which aims to provide interpretable explanations for model decisions and improve transparency, accountability, and clinician trust [16]. However, this objective is complicated by the trust paradox: while explanations are designed to build confidence, they can be manipulated or distorted by adversarial inputs, potentially eroding trust. Moreover, the trust paradox encompasses situations in which individuals’ willingness to use AI-enabled technologies exceeds their actual trust in these systems [17]. Consequently, the effective-

ness of XAI depends not only on the clarity and quality of explanations but also on their resilience to misleading or malicious perturbations. However, effective clinical use requires calibrated trust, whereby clinicians understand both the strengths and limitations of AI systems rather than relying on them uncritically [18].

This paper presents a comprehensive narrative review of research on explainable AI for skin lesion classification. While existing reviews in the literature, including [19] have typically focused either on the diagnostic performance of machine learning and deep learning models in skin cancer detection or general applications of explainable AI in medical imaging, they often do not provide an integrated analysis of how datasets, model architectures, and explainability techniques interact within dermatological applications. As a result, comparatively little attention has been given to how explainability methods are operationalised within skin lesion classification systems and how these affect interpretability and clinical translation.

The novel contribution of this review lies in its comprehensive synthesis of three interconnected components: classification architectures, dermatological image datasets, and explainability frameworks. Rather than evaluating predictive performance alone, this review examines how commonly used models, including convolutional neural networks, are paired with explainability methods such as Gradient-weighted Class Activation Mapping (Grad-CAM), SHapley Additive exPlanations (SHAP), and Local Interpretable Model-agnostic Explanations (LIME) to support transparency and clinician trust. Furthermore, the review evaluates methodological practices, validation strategies, and reported limitations, including dataset bias and the widespread reliance on single-dataset validation. A further contribution of this work is the application of a GRADE-inspired narrative assessment framework to evaluate the quality and strength of evidence within the XAI literature, providing a structured appraisal that is often absent from predominantly technical reviews. By mapping current research trends, assessing evidence quality, and identifying gaps in explainability, robustness, and clinical applicability, this comprehensive narrative review provides a consolidated evidence base and highlights priorities for the development of transparent and clinically reliable CAD systems for skin cancer detection.

2. Materials and Methods

2.1. Research Methodology

This paper presents a comprehensive review of the literature on explainable AI for skin disease classification, aiming to map the current state of research, identify key themes, and highlight gaps in the existing evidence. The review followed a structured six-step comprehensive narrative review methodology to ensure rigour, transparency, and reproducibility as seen in Table 1.

Table 1. Six-Step Methodology.

Index	Step
1. Formulation of Research Questions	Clear research questions were established to guide the scope and focus of the review.
2. Search Strategy	A systematic search was conducted across relevant databases and platforms to identify studies related to skin disease detection, classification, and monitoring
3. Study Screening and Selection	Identified studies were screened based on inclusion and exclusion criteria to ensure relevance and quality.

Table 1. Cont.

Index	Step
4. Data Extraction	Key data and findings were systematically extracted from the selected studies for analysis.
5. Discussion	The extracted data were synthesised and critically analysed to identify patterns, technological trends, and research gaps.
6. Conclusion and Recommendations	Final insights were drawn, highlighting future research directions and practical recommendations for the development of an explainable AI model tailored for medical diagnosis, integrating clinical metadata for enhanced interpretability, and application of hybrid AI techniques to improve diagnostic accuracy in primary care and the creation of a mobile AI-driven diagnostic tool, bridging the gap between AI research and healthcare practice.

2.2. Research Questions

The development of the research questions in this study was informed by the framework proposed by Farrugia et al. [20], which applies the FINER criteria (Feasible, Interesting, Novel, Ethical, and Relevant) to guide the formulation of high-quality research questions. This framework emphasises that research questions should be clearly defined, clinically relevant, feasible within the scope of the study, and capable of contributing to existing knowledge in the field while adhering to ethical and methodological standards [20].

Accordingly, the research questions for this comprehensive narrative review were designed to ensure relevance to clinical dermatology and explainable artificial intelligence, with a focus on identifying datasets, explainability techniques, and methodological limitations within skin lesion classification research. The full set of research questions is presented in Table 2.

Table 2. Research Questions and Rationale.

ID	Research Question	Rationale
RQ1	In the context of skin lesion detection, what are the most commonly used imaging modalities and benchmark datasets, and how do they compare in terms of diagnostic accuracy, reliability, and suitability for explainability methods?	Identifying prevalent image acquisition techniques (e.g., dermoscopy, clinical photography, histopathology) and datasets is essential for understanding their role in the development of XAI models. Comparative analysis of these resources can highlight strengths, weaknesses, and compatibility with interpretability techniques, guiding future dataset curation and system design
RQ2	What computational approaches are employed to incorporate explainability into skin lesion classification models, and how do these methods differ in interpretability and clinical applicability?	This question examines the range of XAI techniques, such as Grad-CAM, LIME, SHAP, and prototype-based networks, used to enhance transparency in AI-based skin lesion classification. Understanding their comparative effectiveness can inform best practices and identify promising directions for innovation.
RQ3	What technological, methodological, and application-level limitations and research gaps are reported in current literature on explainable AI for skin lesion detection?	Identifying existing shortcomings such as limited dataset diversity, absence of standardised XAI evaluation metrics, or insufficient clinical validation can provide a roadmap for addressing critical barriers to adoption in dermatology.

2.3. Search Strategy

A comprehensive literature search was undertaken across multiple academic databases, including IEEE Xplore, Science Direct, Scopus, PubMed, and Google Scholar, to identify relevant research articles pertinent to the focus of this study. The search strategy combined key concepts related to “explainable artificial intelligence”, “skin lesion classification”, and “dermatological imaging” using Boolean operators (AND/OR). Equivalent keyword combinations were adapted for use across the academic databases to ensure coverage of relevant literature. This approach ensured the inclusion of studies published within the last five years, aligning with the temporal scope of this review. The selected search terms were specifically tailored to capture literature at the intersection of skin lesion detection and explainable artificial intelligence, as outlined in Table 3.

Table 3. List of Key Search Words.

Index	Key Word
SK1	Explainable AI or Interpretable AI
SK2	Explainability or Interpretability
SK3	Skin Lesion or Melanoma or Dermatology or Dermoscopy.
SK4	Experiment or Clinical Study or Validation or Performance Assessment or Case Study
SK5	Machine Learning or Deep Learning

2.4. Study Screening and Selection

The screening and selection process for identifying relevant literature on skin cancer, dermatological conditions, machine learning, and deep learning adhered to the PRISMA (Preferred Reporting Items for Systematic Reviews and Meta-Analyses) guidelines [21]. This structured approach facilitated a transparent and reproducible methodology, enabling the inclusion of the most pertinent studies by clearly outlining the search strategy, inclusion criteria, and rationale. As such, PRISMA served as an effective framework for filtering high-quality and contextually appropriate papers from the selected databases, by utilising such databases, we ensured that preprints publications were excluded from the review. In this procedure alone, more than 750 entries were found in the databases, and the PRISMA approach narrowed that number down to just 113 relevant papers, as seen in both Table 4 and Figure 1. The selection process was primarily conducted by the lead author according to the predefined inclusion and exclusion criteria; The final set of included studies was subsequently reviewed by all authors to ensure consistency with the review objectives and eligibility criteria. Any uncertainties regarding study inclusion were resolved through discussion and consensus.

Table 4. Screening Process.

Process	Count
Records identified from database	757
Duplicates removed	107
Studies retracted	3
Records Screened (title/abstract)	647
Records Excluded (title/abstract)	534
Full-text articles sought for retrieval	113

Table 4. Cont.

Process	Count
Reports not retrieved	52
Full-text articles assessed	61
Full-text articles excluded	8
Studied included	53

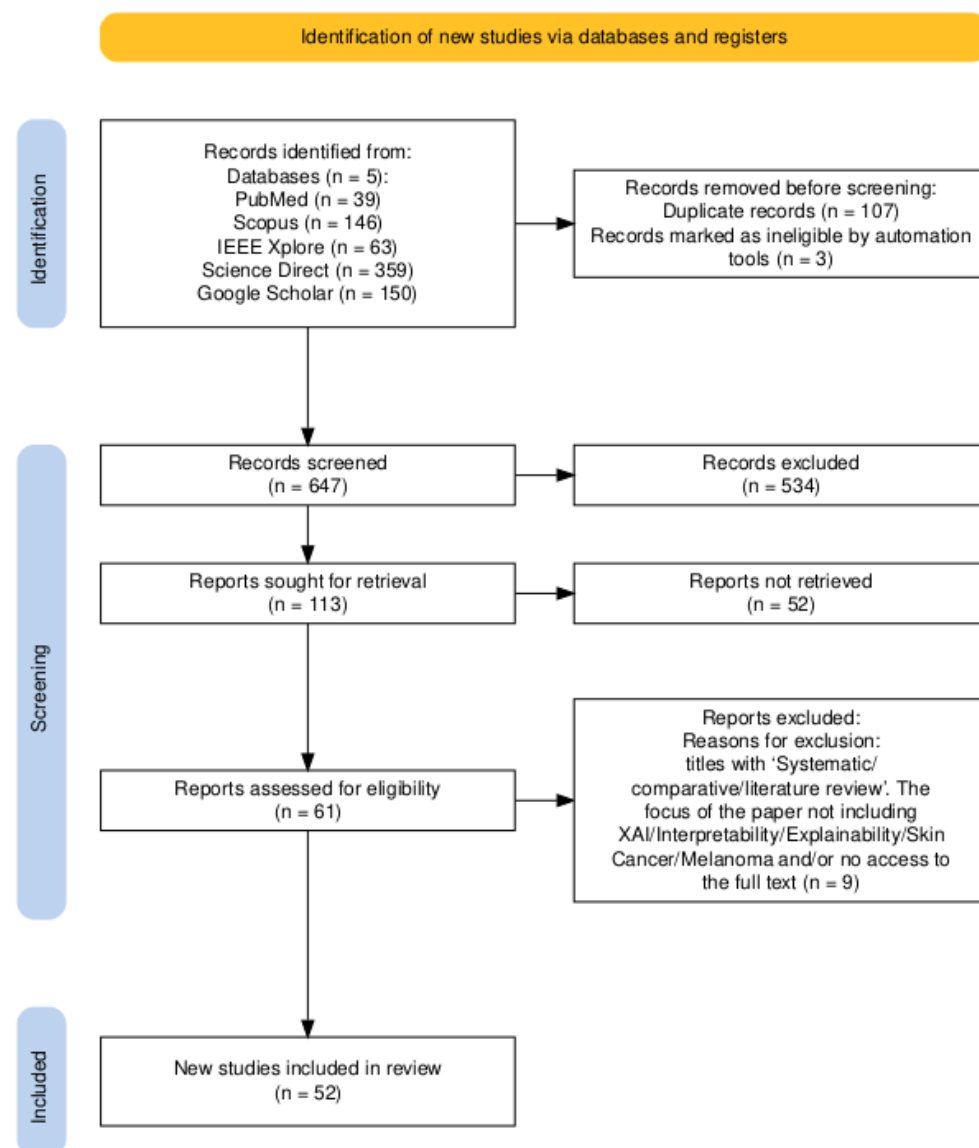


Figure 1. PRISMA Flow Diagram [22].

3. Skin Cancer and Diagnostic Challenges

Skin cancer encompasses a broad spectrum of subtypes, each presenting unique challenges that complicate timely and accurate diagnosis. The three primary classifications of skin cancer are melanoma, basal cell carcinoma (BCC), and squamous cell carcinoma (SCC) [23]. Among these, melanoma is recognised as the most lethal form [24] due to its aggressive nature and potential for metastasis.

Diagnosing suspected malignant melanoma typically involves evaluating specific clinical criteria. As outlined by Walter et al. [25], the assessment includes three major signs—changes in size, shape, and/or colour—and four minor signs: inflammation, crust-

ing or bleeding, sensory alterations, and a lesion diameter of ≥ 7 mm. A weighted scoring system is employed, assigning two points to each major sign and one point to each minor sign; lesions scoring three or more points are recommended for clinical referral.

The other two major subtypes, BCC and SCC, are collectively referred to as non-melanoma skin cancers [26]. These forms differ from melanoma not only in terms of severity but also in their origins. While melanoma typically develops from existing moles or new lesions, BCC and SCC arise from different skin cells and are primarily linked to cumulative sun exposure. Basal cell carcinoma, the most common form of skin cancer, accounts for approximately 75% of all non-melanoma skin cancers. It often appears on sun-exposed areas of the body, particularly the face, including the nose, forehead, and cheeks. Several subtypes of BCC exist, including nodular, superficial, pigmented, and morphoeic (also known as sclerosing or infiltrating) basal cell carcinomas.

Squamous cell carcinoma, on the other hand, originates in keratinocytes—the predominant cell type in the epidermis. Like BCC, SCC typically manifests in sun-exposed regions such as the head, neck, and dorsal hands and forearms.

Despite advances in dermatological assessment, diagnosing skin cancer remains fraught with challenges. These include lesion variability, similarity to benign conditions, limitations in diagnostic methods, and inconsistent access to care. Variability in lesions poses a particular issue in melanoma, which can resemble benign conditions like blood blisters, haemangiomas, nevi, seborrhoeic keratoses, and solar lentigines [27]. Specifically, seborrhoeic keratosis is typically the hardest to diagnose, mimicking melanoma and producing ‘false positive overdiagnosis’, leading to needless excisions [28]. The management of potentially life threatening conditions such as melanoma is guided by a precautionary approach. This approach emphasises minimising the likelihood of missed diagnoses, even at the expense of increased false positives. Consequently, clinicians adopting the low threshold with a score of 3 ensures that any lesion with even a small degree of suspicion is examined further by a specialist.

General practitioners often rely on the ABCDE rule—an evolution of the earlier ABCD criteria introduced in 2004 to assess suspicious lesions. This method involves evaluating lesions for asymmetry, border irregularity, colour variation (including both intra-lesional diversity and colours that differ from a patient’s typical nevi), diameter greater than 6 mm, and evolution, or noticeable changes over time [29]. While useful, this approach can lead to overdiagnosis, triggering unnecessary anxiety, false-positive or false-negative results, and the detection of indolent (non-threatening) cancers [30].

In conclusion, while awareness and clinical guidelines for skin cancer diagnosis have improved significantly, the complexity of its presentation continues to demand nuanced, accurate assessment methods and improved diagnostic support for general practitioners.

4. Techniques for Skin Cancer Detection

4.1. Machine and Deep Learning in Medical Imaging

At present, a wide spectrum of machine learning and deep learning methodologies is employed in the development of skin cancer detection systems, encompassing both traditional algorithms and contemporary explainable AI frameworks. This section provides a critical examination of the computational approaches integrated into these systems for the automated analysis and classification of skin lesions. Figure 2 presents an overview of the algorithmic categories and explainability techniques commonly utilised in skin cancer classification, with these patterns further reflected in the studies summarised in Table 5.

Table 5. Table of shortlisted studies.

Study	Publication Year	Dataset/Source	Classification Method	Results	XAI Framework	XAI Method	Decision Making
[31]	2022	ISIC 2019	ResNet-18	Accuracy: 94.47%	LIME	Post hoc Explainability	Visual Explanation-Interpretable Surrogate Model
[13]	2025	HAM10000 and ISIC 2019	QuadRes-Net and PentRes-Net	Accuracy: 83.9–98.51%	Grad-CAM	Post hoc Explainability	Visual Explanation-Heatmap
[32]	2024	HAM1000	CNN, CTL, TA, EL	Accuracy: 94.93%	Grad-CAM	Post hoc Explainability	Visual Explanation-Heatmap
[33]	2025	HAM1000	RegNet	Accuracy: 94.08%	Grad-CAM	Post hoc Explainability	Visual Explanation-Heatmap
[34]	2022	HAM10000 and ISIC 2018	ResNet, and NasNet	Accuracy: 90.2–97.1%	Grad-CAM	Post hoc Explainability	Visual Explanation-Heatmap
[35]	2025	HAM10000	CNN, ViT, and YOLOv8	ViT with YOLOv8: Accuracy: 93%	Grad-CAM and SHAP	Post hoc Explainability	Visual Explanation-Heatmap and Pixel-Level Correlation Plots
[36]	2022	GSE45216, GSE98774 and GSE108008	CNN / XGBoost	Accuracy: 91.67–100%	SHAP	Post hoc Explainability	Tabular Explanation-Summary Plots
[37]	2023	HAM10000	CNN	Accuracy: 82%	Grad-CAM, Grad-CAM++	Post hoc Explainability	Visual Explanation-Heatmap
[38]	2023	HAM10000	DenseNet, CBAM	DenseNet Accuracy: 93%	LIME	Post hoc Explainability	Visual Explanation-Feature and Importance Score
[39]	2025	ISBI 2016 and ISBI 2017	VGG16, Xception, DenseNet169, InceptionV3, MobileViT, and InceptionResNetV2	Accuracy: 83.3–92.67%	LIME	Post hoc Explainability	Visual Explanation-Feature Map
[40]	2025	Kaggle	MobileNetV2	Accuracy: 98.48%	Grad-CAM	Post hoc Explainability	Visual Explanation-Heatmap
[41]	2024	HAM10000	ResNet50, InceptionV3, VGG16, VGG19, and MobileNetV2	ResNet50 Accuracy: 99%	LIME	Post hoc Explainability	Visual Explanation-Segmentation Map

Table 5. Cont.

Study	Publication Year	Dataset/Source	Classification Method	Results	XAI Framework	XAI Method	Decision Making
[42]	2025	HAM10000	DCAN-Net	Accuracy: 97.57%	Grad-CAM	Post hoc Explainability	Visual Explanation-Heatmap
[43]	2023	ISIC 2018 and HAM10000	Xception and ShuffleNet (Butterfly Optimisation Algorithm)	Accuracy: 91.5–99.3%	Grad-CAM	Post hoc Explainability	Visual Explanation-Heatmap
[44]	2025	HAM10000, PAD-UFES-20 and ISIC 2018	EfficientNet B0 and EfficientNet B4 (EffiCAT)	Accuracy: 94.48%, ISIC 2018 Dataset: Accuracy: 92.08%	Grad-CAM	Post hoc Explainability	Visual Explanation-Heatmap
[45]	2024	HAM10000	ViT	Accuracy: 83%	Attention, Beyond Intuition and Transition Attention Maps, Attention Rollout, Beyond Attention, Generic	Post hoc Explainability	Visual Explanation-Heatmap
[12]	2025	Clinical and Derm7pt	Multi-modal information-switching network	Accuracy: 90.7%	Grad-CAM	Post hoc Explainability	Visual Explanation-Heatmap
[46]	2024	HAM10000	ResNet50	Accuracy: 97.83%	Grad-CAM	Post hoc Explainability	Visual Explanation-Heatmap
[47]	2025	HAM10000 and Dermnet datasets	Vision Transformers, Swin Transformers, and DinoV2	Accuracy: $96.48 \pm 0.0138\%$	Grad-CAM and SHAP	Post hoc Explainability	Visual Explanation-Heatmap and Pixel-Level Correlation Plots
[48]	2024	HAM10000	ABC Ensemble Model: ViT-B/16, DenseNet121, EfficientNet-B0, MobileNetV2, and ResNet50	Accuracy: 93.26%	Grad-CAM	Post hoc Explainability	Visual Explanation-Heatmap
[49]	2022	ISIC 2019, PH2 and Derm7pt datasets	SEResNeXt	Accuracy: 83.6–100%	ExAID (Explainable AI for Dermatology)	Post hoc Explainability	Visual and Textual Explanations
[50]	2024	ISIC 2018 and HAM10000	Xception	Accuracy: 87.1–98.5%	Grad-CAM, LIME, and Occlusion Sensitivity	Post hoc Explainability	Visual Explanation-Heatmap

Table 5. Cont.

Study	Publication Year	Dataset/Source	Classification Method	Results	XAI Framework	XAI Method	Decision Making
[51]	2025	Kaggle datasets	Xception	Accuracy: 95.23–96.48%	Grad-CAM	Post hoc Explainability	Visual Explanation-Heatmap
[52]	2025	HAM10000, ISIC2019, ISIC2020, and Melanoma Skin Cancer-comprising skin cancer images	EfficientNet, MobileNet, and Darknet	Accuracy: 99.86%	Grad-CAM	Post hoc Explainability	Visual Explanation-Heatmap
[53]	2025	HAM10000 and ISIC2018	5-block inverted residual network and a 6-block inverted bottleneck network	Accuracy: 90.7–91.3%	LIME	Post hoc Explainability	Visual Explanation-Heatmap
[54]	2025	ISIC Archive	ResNet50, VGG19, ViT-b16 and ViT-b32	Accuracy: 89.09%	LIME and Integrated Gradients	Post hoc Explainability	Visual Explanation-Integrated Gradients and Feature Maps
[55]	2025	ISIC 2019	CNN	Accuracy: 97.58%	t-distributed Stochastic Neighbor Embedding (t-SNE) and Grad-CAM	Post hoc Explainability	Visual Explanation-Heatmap
[56]	2025	Kaggle	CNN	Accuracy: 66.25%	Grad-CAM++	Post hoc Explainability	Visual Explanation-Heatmap
[57]	2025	ISIC and HAM10000	CNN	Accuracy: 97–99%	Grad-CAM	Post hoc Explainability	Visual Explanation-Heatmap
[58]	2025	PADUFES-20	DenseNet, ResNet, MobileNet, and EfficientNet	Accuracy: 79.40–85.50%	SHAP, Grad-CAM, Grad-CAM++ and AblationCAM	Post hoc Explainability	Visual Explanation-Heatmap
[59]	2025	HAM10000 and ISIC	Vision Transformer, CNN, and Xception	Accuracy: 93.19–96.74%	Grad-CAM	Post hoc Explainability	Visual Explanation-Heatmap
[60]	2024	HAM10000	Inception v3	Accuracy: 96%	Grad-CAM and LIME	Post hoc Explainability	Visual Explanation-Heatmap and Feature maps
[61]	2024	Skin Cancer: Malignant vs. Benign and HAM10000	CNNs	Accuracy: 96.5–97.2%	LIME	Post hoc Explainability	Visual Explanation-Heatmap
[62]	2024	HAM10000	CNNs and ViT	Accuracy: 91.09–98.37%	Grad-CAM and Grad-CAM++	Post hoc Explainability	Visual Explanation-Heatmap

Table 5. Cont.

Study	Publication Year	Dataset/Source	Classification Method	Results	XAI Framework	XAI Method	Decision Making
[63]	2024	NCI 2022 and NCI 2023	VGG-SVM	Accuracy: 96%	LIME	Post hoc Explainability	Visual Explanation-Segmentation Map
[64]	2024	ISIC 2019	DenseNet and TL	Accuracy: 91.61%	Grad-CAM	Post hoc Explainability	Visual Explanation-Heatmap
[65]	2024	HAM10000	Resnet50, VGGNet, and DenseNet	Accuracy: 67.33–85%	LIME and SHAP	Post hoc Explainability	Visual Explanation-Heatmap and Segmentation Map
[66]	2024	ISIC	ViTs, CNNs, and Swin Transformers	Accuracy: 88.6–88.8%	Grad-CAM, Grad-CAM++ and Score-CAM (Score-Weighted Visual Explanations for Convolutional Neural Networks)	Post hoc Explainability	Visual Explanation-Heatmap
[67]	2024	ISIC-2018 and ISIC-2019	DCNNs	Accuracy: 96.35–96.39%	LIME and SHAP	Post hoc Explainability	Visual Explanation-Positive and Negative region and Feature Importance Plot
[68]	2024	MED-NODE	InceptionV3, Xception, VGG19, EfficientNetB1, and DenseNet201	Accuracy: 95.88–96.47%	LIME	Post hoc Explainability	Visual Explanation-Feature Map
[69]	2024	PAD-UFES-20	ViT, Dino V2 and XGBoost	Accuracy: 85–86%	Grad-CAM	Post hoc Explainability	Visual Explanation-Heatmap
[70]	2023	PH2	CNN	Accuracy: 94%	SHAP and partial dependence plots (PDPs)	Post hoc Explainability	Visual Explanation-Summary Plot and Feature Plots
[71]	2023	HAM10000	Xception	Accuracy: 90.24%	Grad-CAM and Grad-CAM++	Post hoc Explainability	Visual Explanation-Heatmap
[72]	2022	Dermnet NZ, DermIS, Dermatology Atlas, Global Skin Atlas, SJS Awareness UK, Dermnet, and University of IOWA Health Care	ResNet50V2, VGG16, InceptionV3, and InceptionResNetV2	Accuracy: 90.27–99.06%	LIME	Post hoc Explainability	Visual Explanation-Heatmap and Feature Maps

Table 5. Cont.

Study	Publication Year	Dataset/Source	Classification Method	Results	XAI Framework	XAI Method	Decision Making
[73]	2021	ISIC	ResNet50, VGG16, MobileNetV2, NasNetMobile, InceptionV2, Xception, InceptionResNetV2, DenseNet201, EfficientNet3 and 5	Accuracy: 85.8–89.6%	Totally eXplainable Model (TXM)	-	Visual and Textual Explanations
[11]	2021	ISIC	ResNet	Accuracy: 83.8%	ABELE	Post hoc Explainability	Visual and Textual Explanations
[74]	2021	ISIC and PH2	IVF-ResNet-50	Accuracy: 99.23%	CAM	Post hoc Explainability	Visual Explanation-Heatmap and Relevance Maps
[75]	2021	HAM10000	VGG19, ResNet50, DenseNet121, and EfficientNetB0	Accuracy: 86.7–91.45%	Grad-CAM	Post hoc Explainability	Visual Explanation-Heatmap
[76]	2021	HAM10000	DenseNet121, DenseNet169, Inception ResNet v2, Resnet50, Inception v3, VGG16, VGG19, and Xception	Accuracy: 90–95.10%	Grad-CAM and SHAP	Post hoc Explainability	Visual Explanation-Heatmap and Summary Plot
[77]	2020	PH2 and derm7pt	RECOD	Accuracy: 50–80%	CAVs	Post hoc Explainability	Textual Explanations
[78]	2021	ISIC	Vgg16 and Resnet50	Accuracy: 72.2–76.6%	Grad-CAM	Post hoc Explainability	Visual Explanation-Heatmap
[79]	2021	HAM10000	MobileNet	Accuracy: 74%	LIME	Post hoc Explainability	Visual Explanation-Heatmap
[80]	2020	HAM10000	VGG-16, DenseNet, Xception, and Inception ResNet v2	Accuracy: 53.26–85.69%	LIME	Post hoc Explainability	Visual Explanation-Positive and Negative region and Feature Importance Plot

Methods highlighted in green denote approaches that are inherently transparent, with decision-making processes that can be directly traced to the underlying input features. Those highlighted in orange indicate models with moderate interpretability, approaches that are not intrinsically transparent but can be elucidated through external explanation mechanisms. Finally, methods shown in red represent “black-box” models, which offer minimal insight into their internal reasoning processes without the application of specialised interpretability techniques.

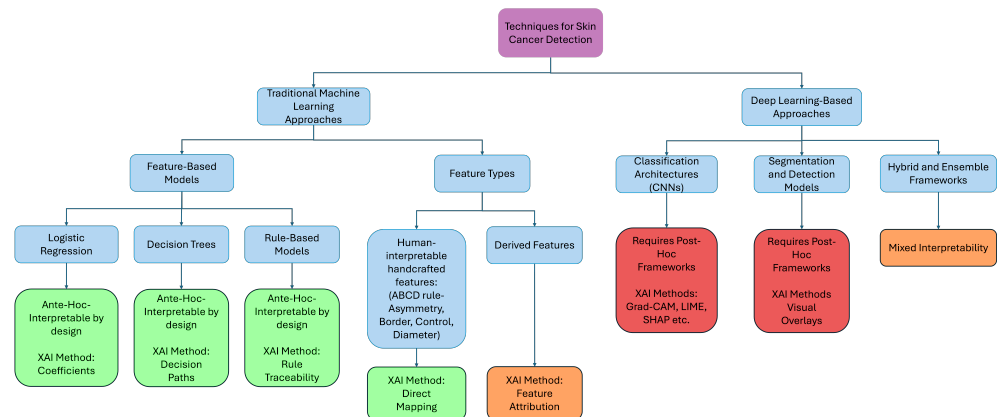


Figure 2. Taxonomy of Algorithms for Skin Cancer Detection.

One of the primary challenges in the diagnosis of skin cancer may be addressed through advancements in artificial intelligence (AI). AI is broadly defined as the capability of a computer system or a computer-controlled robot to perform tasks typically requiring human intelligence [81]. AI encompasses a range of paradigms, including artificial general intelligence, narrow AI, reactive machines, limited memory AI, theory of mind, self-aware AI, and artificial superintelligence [82]. Within the context of medical imaging, particularly in dermatology, research has predominantly focused on reactive machine learning systems. Machine learning (ML) and deep learning (DL) are closely related subfields of AI. ML refers to the ability of systems to learn from domain-specific training data to autonomously generate analytical models and solve relevant problems. In contrast, DL extends this concept by leveraging artificial neural networks, allowing for the construction of models that incorporate both structured data and unstructured data, such as medical images, to enhance diagnostic accuracy [83].

In the domain of medical imaging with machine learning and deep learning techniques, CAD systems are predominantly developed using a range of deep learning architectures. Among these, Convolutional Neural Networks (CNN) are the most widely adopted due to their proven efficacy in tasks such as classification, detection, and segmentation of single lesions. CNNs have demonstrated particularly strong performance in identifying abnormalities such as liver and kidney tumours, nasopharyngeal lesions, and other localised pathologies [84].

4.2. Explainable Artificial Intelligence

While machine learning and deep learning techniques continue to deliver highly accurate results, these outputs alone may hold limited value for users who lack an understanding of how the models function or how the results are derived. This challenge is commonly referred to as the “black box” problem in artificial intelligence. Although such complex models are capable of producing impressive outcomes, the opacity of their internal mechanisms often makes the results difficult to interpret and trust [85]. Without insight into the model’s decision-making processes, users are unable to verify or validate the outputs effectively. This lack of transparency has posed a significant barrier to the

widespread adoption of machine learning systems in critical and high-stakes domains such as healthcare, where interpretability and trust are essential [86].

In response to this issue, the concept of XAI has emerged as a growing field of research. XAI seeks to address the limitations of opaque models by making their internal processes more transparent and their outputs more interpretable. By providing meaningful explanations, XAI aims to enhance user trust, promote accountability, and support more informed decision-making in applications where reliability is paramount. The question then arises: how can AI systems be made explainable? Broadly, there are two principal approaches to achieving explainability in AI: intrinsic vs. post hoc methodologies and local vs. global explanations.

Intrinsic interpretability refers to models that are inherently understandable due to their design; such models are typically built using techniques like regression or decision trees, which provide transparency in their structure and decision-making processes [87]. These models are designed with an interpretable structure from the outset, meaning explainability is embedded directly into the model's architecture and is available at the end of the learning phase. The capacity to describe "how the model works" is an advantage of intrinsic models. Nevertheless, because of the trade-off with interpretability, they might occasionally show reduced accuracy.

In contrast, post hoc interpretability involves analysing a model after it has been trained, aiming to elucidate how it arrived at specific predictions [88]. This approach seeks to generate explanations for otherwise opaque or "black box" models, such as deep neural networks, thereby enhancing the interpretability of their outputs without altering the original model architecture. The majority of current interpretability techniques fall into this category, with post hoc methods being applicable to a variety of scenarios and levels of access to model parameters.

The alternative method examines local vs global explanations. A single, particular prediction or experience is given an interpretation via local explanations. For instance, describing the reasons for the classification of a specific skin lesion image as malignant. Global explanations, on the other hand, seek to explain the model's general logic or how a specific model influences predictions for the entire dataset. For instance, determining which general characteristics—like lesion asymmetry—are most crucial to the model's overall diagnostic performance [89].

Several established methods in the field of XAI have been developed to make AI model decision-making processes more interpretable and understandable to humans. These methods are often model-specific, meaning they are tailored to particular machine learning or deep learning architectures. Among the most widely used techniques are Grad-CAM, LIME, and SHAP. Grad-CAM is primarily employed for generating visual explanations in convolutional neural networks. It works by highlighting important regions in an input image that contribute most significantly to the model's prediction, thereby offering a spatial representation of feature importance [90]. SHAP values provide a unified approach to interpreting model outputs by assigning an importance value to each feature based on its contribution to a specific prediction [91]. This method is grounded in cooperative game theory, wherein the original concept of Shapley values was used to determine the contribution of individual players to the overall gain of a coalition. Similarly, SHAP applies this principle to determine how each feature contributes to the model's prediction [92]. LIME is another widely adopted XAI method that explains individual predictions by approximating the model locally with an interpretable one. In classification tasks, such as logistic regression, LIME provides probabilities indicating the likelihood of a given instance belonging to each class while also highlighting the influence of each feature on that prediction [93].

But how do we evaluate that the explainable methods are reliable? While interpretability and explainability are often presented as a subjectively validated binary property, we consider it a multi-faceted concept [94]. This is where the need for evaluation metrics comes into play; metrics such as faithfulness, localisation and human-centred measurements are just some of the metrics utilised for evaluating explainable AI. Faithfulness of explanations correlation measures the linear relationship between the model predictions and the training data. It quantifies how well the model can capture the patterns and relationships in the training data [95].

5. Discussion

5.1. Classification Algorithms

Figure 3 illustrates the relationship between average classification accuracy, dataset size, and the number of classes considered across the reviewed studies. The distribution of data points reveals substantial heterogeneity within the literature, reflecting variation not only in dataset size but also in class granularity, dataset composition, and the diversity of algorithms employed. These factors interact in complex ways, making it difficult to infer simple causal relationships between individual variables and reported performance. In particular, differences in class definitions, dataset curation practices, and evaluation protocols introduce significant variability that complicates direct comparisons across studies.

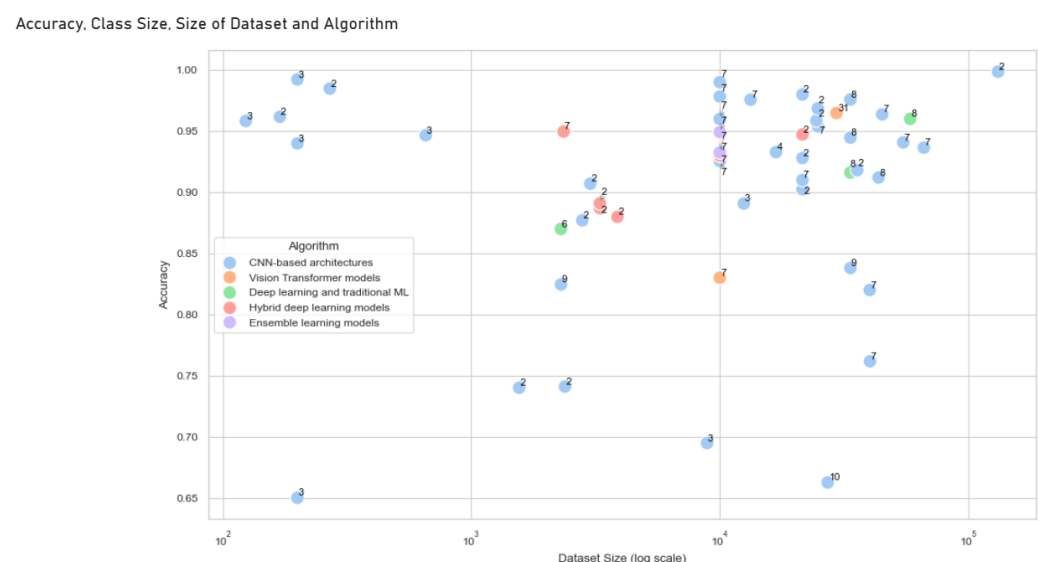


Figure 3. Distribution of algorithms by class group, dataset size, and reported accuracy. The class group represents the number of skin lesion categories included in the classification task. A class group of 2 corresponds to binary classification (e.g., malignant versus benign), while a class group of 3 corresponds to three-class classification (e.g., melanoma, nevi, and other lesions). Class groups of 4, 6, 7, 8, 9, 10 and 31 indicate multiclass classification tasks involving the corresponding number of distinct skin lesion categories. The lesion categories vary according to the dataset used and may include melanoma, melanocytic nevi, basal cell carcinoma, benign keratosis, actinic keratosis, vascular lesions, dermatofibroma, squamous cell carcinoma, and other clinically relevant lesion types.

While the Figure 3 does not indicate a straightforward linear relationship between dataset size and classification accuracy, a more nuanced pattern emerges when the number of classes and the methodological approaches employed are considered simultaneously. Studies that utilise larger datasets in conjunction with a greater number of diagnostic classes frequently report higher levels of classification accuracy. However, this relationship should be interpreted cautiously, as improvements in reported performance may also reflect the influence of dataset structure, class balance, and the availability of well-curated

benchmark datasets rather than purely algorithmic improvements. Nevertheless, larger datasets with more granular class structures may provide more comprehensive training signals, which can help mitigate over-fitting and support improved generalisation across diverse lesion types.

The role of algorithmic design further complicates this relationship. CNN-based architectures remain the most prevalent methodological approach within the literature, as reflected by the colour distribution of the bubbles in Figure 3. These architectures form the backbone of many skin lesion classification pipelines due to their effectiveness in extracting hierarchical visual features from dermoscopic images. However, the figure also suggests that studies employing more complex modelling strategies such as ensemble frameworks or hybrid deep learning architectures are more frequently associated with larger datasets and multi-dataset evaluation settings. In these contexts, ensemble and hybrid approaches appear to demonstrate improved robustness, likely due to their capacity to aggregate complementary representations and mitigate model-specific biases when trained on heterogeneous data sources.

Taken together, the results illustrated in Figure 3 suggest that classification performance in skin lesion analysis cannot be explained by dataset size alone. Instead, performance outcomes appear to emerge from the interaction of several interdependent factors, including dataset scale, class complexity, algorithmic architecture, and the diversity of training data. Consequently, studies that simultaneously incorporate larger and more heterogeneous datasets, finer-grained class definitions, and more sophisticated modelling approaches may provide a more realistic assessment of model generalisability.

The patterns observed in Figure 3 are further reinforced by the analyses presented in Figures 4 and 5. In particular, Figure 4 represents that the Human Against Machine with 10,000 training images (HAM10000) and International Skin Imaging Collaboration (ISIC) datasets are shown to dominate the experimental landscape, with many studies relying heavily on these benchmarks for both training and evaluation. The studies included in this review utilised a range of dermatological imaging modalities, which play a critical role in determining model performance and generalisability. The majority of datasets were based on dermoscopic imaging, particularly large benchmark datasets such as HAM10000 and ISIC, which are widely used for training and evaluating skin lesion classification models. Dermoscopic images provide magnified, high-resolution views of lesion structures, enabling models to learn fine-grained morphological features relevant to diagnosis.

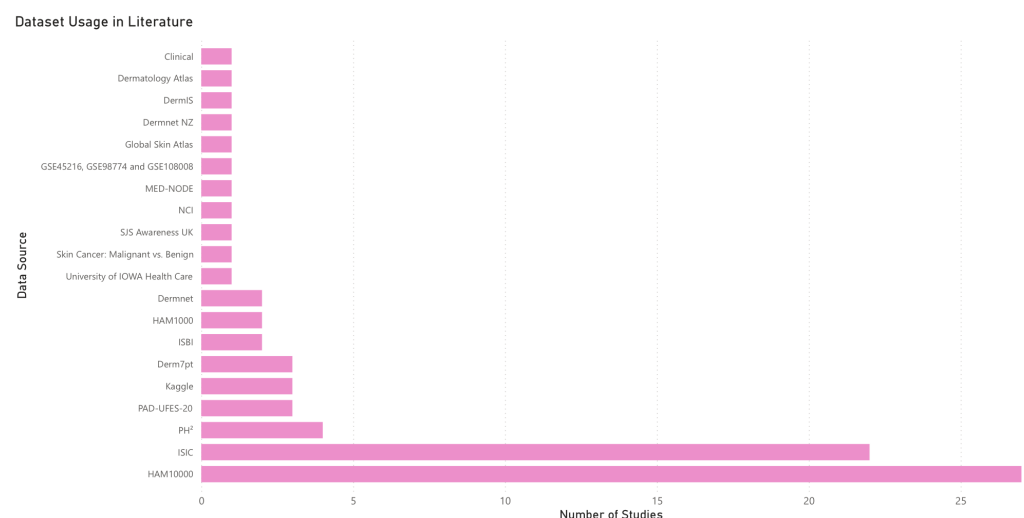


Figure 4. Dataset Usage in Literature.

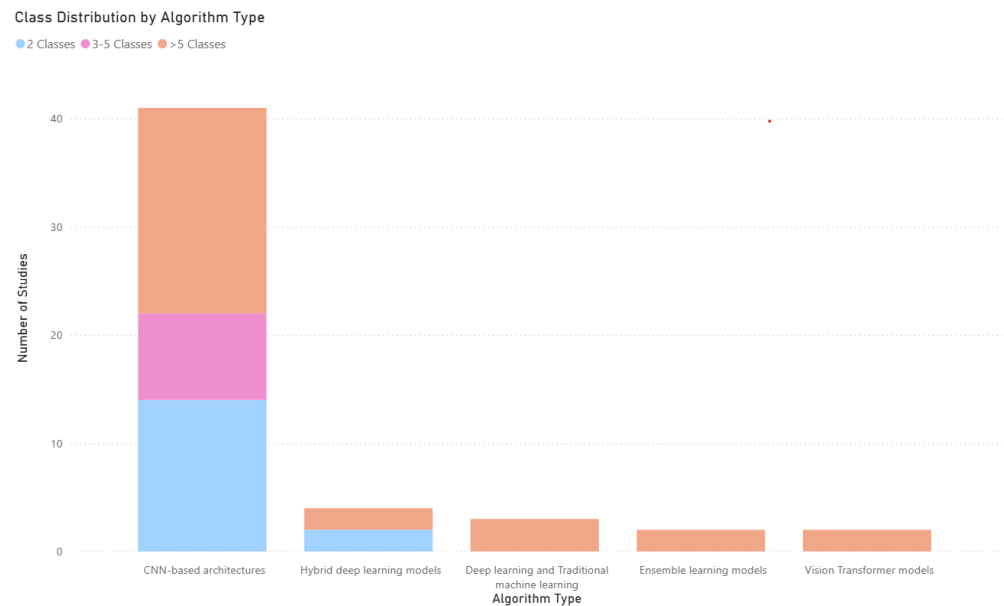


Figure 5. Distribution of algorithms by class group. Colours indicate the number of skin lesion categories included in the classification task. Blue represents binary classification involving two classes (e.g., malignant versus benign), pink represents multiclass classification involving three to five lesion categories (e.g., melanoma, nevi, and other lesion types), and orange represents classification tasks involving more than five distinct lesion categories. The specific lesion types vary across datasets and studies.

In contrast, a smaller subset of studies employed clinical (non-dermoscopic) images captured under real-world conditions, which introduce greater variability in lighting, background, and image quality. These datasets are generally more representative of routine primary care settings but are less commonly used due to their increased complexity. Only a limited number of studies incorporated mixed datasets or multimodal inputs combining dermoscopic images with clinical metadata or histopathological data.

This variability in data type has important implications for model performance, particularly with respect to generalisability and explainability, as models trained on highly curated dermoscopic datasets may not transfer effectively to real-world clinical environments. As these datasets typically support classification tasks involving five or more lesion categories, many reported studies adopt similar class configurations, which may inadvertently shape methodological choices and evaluation outcomes. While the widespread use of these datasets facilitates comparability across studies, it may also introduce a degree of dataset bias, whereby algorithmic performance becomes optimised for specific benchmark characteristics rather than broader clinical variability.

In addition to dataset usage patterns, Figure 5 also highlights a strong reliance on CNN-based architectures, which are present in 41 of the 53 included studies. This dominance reflects the continued centrality of CNNs in dermatological image analysis and may further contribute to methodological homogeneity across the field.

Overall, these observations suggest that the field is gradually transitioning from narrowly optimised single-dataset evaluations toward more comprehensive experimental designs that prioritise robustness and cross-dataset generalisation. However, the continued reliance on a limited number of benchmark datasets and CNN-based architectures highlights the need for more diverse evaluation frameworks that better reflect the variability encountered in real-world dermatological practice.

5.2. Explainable AI Models

The literature demonstrates a discernible progression in the application of explainable AI methods across studies published between 2020 and 2025. Early contributions, particularly in 2020, were characterised by a reliance on CA-Net, Concept Activation Vectors (CAVs), and LIME, which built upon existing machine learning and deep learning approaches to provide explainable artificial intelligence frameworks. However, these initial approaches were limited in their ability to produce clinically meaningful explanations. For instance, LIME's perturbation-based approach often fails to capture the spatial dependencies present in dermoscopic imagery, leading to explanations that may not align with dermatological reasoning, ultimately falling into the explainability trap, where post hoc explanations create an illusion of understanding without providing genuine insight into model behaviour, potentially leading to decreased clinical performance and accuracy, compromising patient safety [15].

From 2021 onwards, there was a notable consolidation of methodological approaches within the literature: Grad-CAM, LIME, and SHAP emerged as the predominant frameworks, establishing the foundation for explainability research in skin lesion and skin cancer detection. While this consolidation reflects methodological maturity, it also suggests a potential over reliance on a narrow set of post hoc tools, many of which were not originally designed for medical imaging contexts. Grad-CAM, for example, offers intuitive visual explanations but has been criticised for producing heat maps that lack quantitative rigour and do not consistently correspond to diagnostically relevant regions. Furthermore, Grad-CAM can be unstable and sensitive to minor perturbations in the input image. Grad-CAM primarily captures the importance of high-level features rather than their spatial relationships; the resulting class-specific heat maps are often unsuitable for object detection tasks [96].

Between 2022 and 2023, the field began to exhibit greater balance in the literature. Grad-CAM retained its prominence, yet it was no longer the sole focus of inquiry, with LIME and SHAP sustaining strong relevance and appearing consistently in conjunction with it. This period signals a phase of gradual diversification, as researchers explored alternative approaches while continuing to rely on the core methods that had defined earlier work. Notably, SHAP's model-agnostic design introduced a more interpretable link between features and outcomes; however, its computational cost and instability across model runs limit its practical use in large-scale clinical settings [97].

In 2024, and more distinctly into 2025, the landscape becomes increasingly heterogeneous. While Grad-CAM and LIME remain central, additional techniques such as Grad-CAM++, Occlusion Sensitivity, Score-CAM, Saliency Maps, and t-SNE feature more frequently in the literature. The adoption of these methods suggests a growing inclination among researchers to experiment with specialised or refined frameworks, moving beyond the foundational trio.

This methodological broadening may also reflect a recognition that no single XAI method provides a holistic explanation of model behaviour; Grad-CAM offers localisation, SHAP provides feature-level attribution, and LIME yields model-agnostic interpretability, yet each captures only one dimension of the interpretability problem.

This diversification reflects an ongoing endeavour to capture different dimensions of explainability and to address domain-specific challenges more effectively. Overall, the trend illustrated in Figure 6 portrays a field in transition. What began as a reliance on a small set of widely adopted tools has evolved into a more diverse methodological ecosystem. The sustained presence of Grad-CAM, LIME, and SHAP underscores their enduring value, yet the increasing prominence of emerging techniques points towards a future of continued innovation, where no single framework is likely to maintain unchallenged dominance.

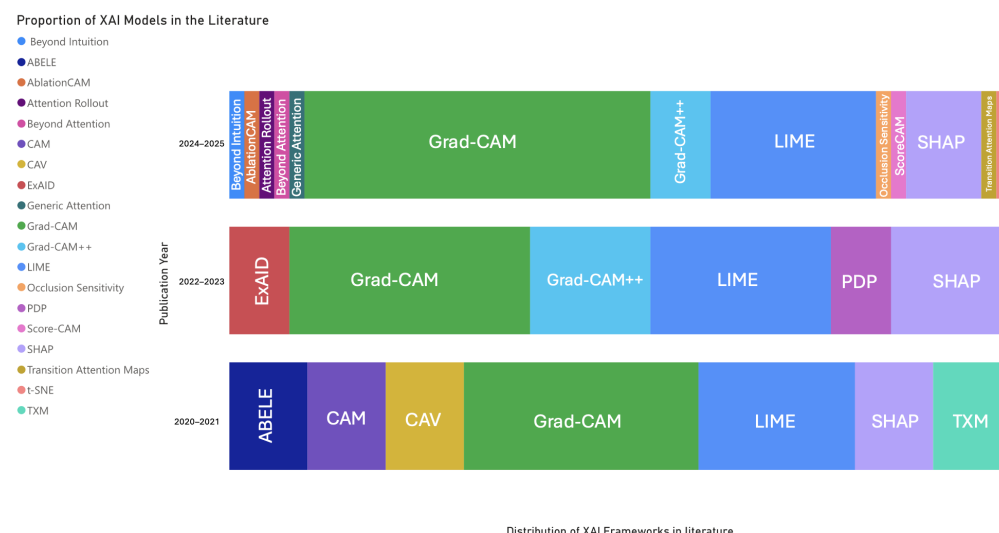


Figure 6. XAI frameworks utilised over time, as reported in the literature between 2020 and 2025.

5.3. Qualitative Comparative Assessment of XAI for Skin Cancer Detection

While numerous explainable artificial intelligence techniques have been proposed for skin cancer detection, their strengths and limitations are often discussed in isolation or evaluated using heterogeneous metrics. This makes it difficult to obtain a coherent comparative view of how different methods perform with respect to clinically relevant explainability requirements.

Grad-CAM, Grad-CAM++, CAM, SHAP, and LIME share a common characteristic in that they provide visual explanations of model predictions by generating heat-maps that highlight regions of the lesion most influential in the decision-making process. These techniques are effective in localising lesions and illustrating their correspondence with patterns learned during model training. However, while such visualisations indicate where the model is focusing, they offer limited insight into how or why a lesion is classified as malignant or benign. The absence of deeper, textual explanations in many existing studies restricts their clinical applicability, where transparent and interpretable reasoning is essential for decision support.

In contrast, ABELE (Adversarial black-box Explainer generating Latent Exemplars) [11] extends beyond purely visual explanations by incorporating textual information alongside visual outputs. Despite this advantage, the method presents certain limitations. ABELE generates multiple potential explanatory conditions and assigns neighbourhood scores to each, with the final prediction based on the smallest neighbourhood score. Although the inclusion of supporting conditions may enhance interpretability, the presentation of multiple competing explanations can introduce ambiguity, potentially leading to clinician confusion and reduced calibrated trust in the system.

CAVs [49] adopt a fundamentally different approach from traditional XAI techniques. These methods typically provide textual explanations linked to high-level, human-interpretable concepts, supplemented by contextual metadata such as patient age, sex, ethnicity, lesion location, and classifier outputs. While this enriched contextual information may offer stronger evidential support for clinical decision-making, the underlying classification mechanisms remain insufficiently explained, and there is a risk that the provided information may be incomplete or misleading.

For this evaluation, explainability techniques were evaluated using a GRADE-inspired framework comprising the criteria of robustness, plausibility, fidelity, usefulness, and completeness as seen in Table 6. As the name suggests, this framework grades methodologies

across four levels of evidential confidence: high, moderate, low, and very low [98]. These criteria were derived from established XAI evaluation taxonomies and adapted to the specific requirements of dermatological decision support, where stability, clinical coherence, and interpretability are essential for safe adoption. The explainability techniques considered in this assessment are illustrated in Figure 6, which categorises the principal XAI approaches identified in the reviewed literature. Table 7 then applies the proposed evaluation criteria to these methods, providing a structured comparison of their relative strengths and limitations in dermatological applications based on the evidence identified in this review.

Robustness refers to the stability and consistency of explanations under controlled perturbations. In the context of dermoscopic imaging, this includes resilience to minor input variations (e.g., noise), model stochasticity (e.g., retraining, weight initialisation), or minor preprocessing adjustments. An explanation method is considered robust if semantically meaningful regions and explanatory patterns remain stable under such perturbations. Given the variability inherent in real-world dermatology imaging, robustness is critical to prevent explanation drift that could undermine clinical trust [99].

Plausibility concerns the degree to which explanations align with established dermatological knowledge and diagnostic reasoning. In melanoma assessment, this includes correspondence with recognised criteria such as the ABCDE rule, dermoscopic structures (e.g., pigment network, streaks, blue-white veil), and other clinically validated morphological cues. An explanation may appear visually convincing yet lack alignment with domain expertise; therefore, plausibility captures the extent to which the explanation reflects medically meaningful features rather than artefactual or dataset-specific patterns [100].

Fidelity evaluates whether the explanation accurately represents the model's true internal decision-making process rather than providing a post hoc approximation that is only loosely correlated with the prediction. High-fidelity explanations demonstrate causal relevance removal or modification of highlighted regions should materially affect the model's output. In dermatological AI systems, low fidelity is particularly concerning because visually appealing saliency maps may misrepresent the features driving classification, thereby creating a false sense of transparency [101].

Usefulness assesses the practical value of the explanation within a clinical workflow. This includes interpretability by dermatologists, integration into existing diagnostic routines, support for second-opinion validation, and the ability to identify bias, confounding factors, or spurious correlations (e.g., ruler marks, colour calibration charts, or acquisition artefacts). Usefulness also encompasses whether the explanation facilitates appropriate trust calibration; neither overconfidence nor unwarranted scepticism and whether it meaningfully supports clinical decision-making rather than serving as a purely illustrative visualisation.

Completeness refers to the extent to which the explanation captures the full range of relevant factors influencing a prediction. In multimodal dermatological assessment, this may include not only pixel-level image features but also higher-level morphological concepts and, where applicable, non-image metadata (e.g., patient age, lesion location, prior history). Methods that focus narrowly on local pixel saliency risk presenting an overly partial view of the model's reasoning. A complete explanation should therefore account for both spatial evidence and concept-level contributions, ensuring that influential features are neither omitted nor disproportionately emphasised [102].

It should be noted that this comparison does not constitute quantitative benchmarking and is inherently influenced by the availability and scope of reported evidence in the literature. Nevertheless, the framework provides a transparent and clinically grounded synthesis that complements experimental evaluations. An important observation emerging

from Table 7 is the predominance of metric-centred evaluation approaches within the dermatological XAI literature. Saliency-based methods such as Grad-CAM, Grad-CAM++, LIME, and SHAP are most frequently assessed using quantitative measures of localisation, fidelity, or perturbation-based robustness. In contrast, concept-based, prototype-based, and rule-based approaches more commonly incorporate human-centred elements, as their explanations are designed to align with clinical reasoning and domain knowledge. Nevertheless, relatively few studies report formal clinician-centred validation, highlighting a persistent gap between technical evaluation and real-world clinical applicability.

Table 6. Scoring Process.

Score	Operational Meaning
Fidelity	Faithfulness of the explanation to the model's true internal decision process.
Plausibility	Alignment of explanations with established dermatological knowledge (e.g., ABCDE, dermoscopic features).
Robustness	Stability of explanations under small input or model perturbations (e.g., noise, retraining, artifacts).
Completeness	Extent to which the explanation captures all relevant predictive features or concepts.
Usefulness	Practical clinical value, including interpretability, bias detection, and support for trust and decision-making.

Table 7. XAI Methods and Performance.

XAI Method	Fidelity (Model Faithfulness)	Plausibility (Clinical Alignment)	Robustness (Stability)	Completeness (Coverage of Factors)	Usefulness (Clinical Utility)	Overall Certainty of Evidence
Grad-CAM/Grad-CAM++	Moderate—Multiple studies report deletion and sanity-check tests, but failures under model randomization and sensitivity to noise are common [78].	Strong—Consistent alignment with lesion regions and dermoscopic structures across datasets [103].	Moderate—Explanation stability degrades under perturbations and retraining [104].	Low—Limited to pixel-level saliency; ignores metadata and higher-level concepts [105].	Moderate—Improves clinician confidence and visual triage but not consistently diagnostic accuracy [62].	Moderate
LIME/SHAP	Moderate—Strong—Surrogate-model fidelity validated in several studies though instability remains [54].	Moderate—Explanations partially correspond to lesion features but are sensitive to segmentation [65].	Low—Moderate—Known instability under small perturbations [106].	Moderate—Can include metadata and tabular features [107].	Moderate—Useful for bias detection and feature auditing [108].	Moderate
TCAV (Concept-Based)	Moderate—Measures causal concept influence but depends on concept quality [109].	Strong—Direct mapping to clinical concepts (e.g., border irregularity, colour) [110].	Moderate—Sensitive to concept dataset selection [11].	Strong—Captures multiple diagnostic concepts [77].	Moderate—Supports clinical reasoning but requires curated concepts [49].	Moderate–High
Prototype-Based Networks	Moderate—Strong—Predictions linked to specific learned examples [111].	Strong—Visual similarity to known lesions is clinically intuitive [112].	Moderate—Stability depends on training diversity [107].	Strong—Multiple lesion patterns and subtypes represented [89].	Strong—Supports GP-level triage and decision justification [89].	High
Rule-Based/Hybrid (e.g., ABELE)	Strong—Explicit logic corresponds directly to model behaviour [113].	Strong—Rules reflect ABCDE-style criteria [114].	Strong—Highly stable once learned [115].	Strong—Can integrate imaging and metadata [11].	Moderate—Interpretable but may limit predictive performance [116].	High

5.4. Explainability Techniques Applied to Skin Cancer Models

Since explainability is still an ever-growing concept, although the healthcare industry has been explored, it has not been heavily applied to skin cancer models. There are two significant models that have provided good results; as such, the first was a study conducted by Lucieri et al. [49] which utilises a newer explainable AI framework known as ExAID (Explainable AI for Dermatology) which is specifically used for biomedical image analysis that provides multi-modal concept-based explanations, consisting of easy-to-understand textual explanations and visual maps, to justify the predictions. The framework utilises CAVs to map human-understandable concepts to those learnt by an arbitrary Deep Learning-based algorithm and Concept Localisation Maps to highlight those concepts in the input space. Overall, there were 6 datasets utilised in the study, ranging from just 40 images in one dataset to 1295 in another. The qualitative results provide somewhat accurate results, with five out of the six sampled lesions indicating a correct result.

It can be observed in the study that the lesion classifier achieved AUCs above 0.85 for five out of six datasets, with two datasets scoring almost perfectly. The dataset from ISIC 2017 achieved a slightly lower AUC of 0.70, which is also reflected in lower precision and recall; interestingly, this dataset utilises 510 images, falling in the median of the dataset range.

In a similar study conducted by Nigar et al. [31], the authors also employed the ISIC 2019 dataset, similar to the study above. They utilised this dataset to develop a skin lesion classification model using ResNet-18, a widely adopted CNN architecture applied via transfer learning. The model was further integrated with the LIME framework to enhance interpretability during the training phase. This approach yielded an accuracy of 94.47% on pre-processed and augmented dermoscopic images, demonstrating the model's ability to accurately identify various skin conditions. Notably, the explainable AI component provided class-specific visual explanations with a reported confidence of 100% for each infected image.

Despite these promising results, the study's limitations are evident. The model was evaluated solely on a single dataset ISIC 2019, raising concerns regarding its generalisability and external validity. Furthermore, the authors relied exclusively on a single pre-trained architecture ResNet-18, without comparison to more advanced or recent deep learning models. Incorporating additional datasets and exploring other state-of-the-art pre-trained networks could enhance the robustness and classification performance of the proposed system.

5.5. Ethical and Legal Hurdles

Explainable AI introduces a range of ethical, legal, and regulatory challenges. Although organisations such as the IEEE and ACM have established comprehensive ethical frameworks to promote responsible AI development and deployment [117], significant uncertainties remain—particularly within clinical contexts such as skin lesion detection. In the United Kingdom, AI-based diagnostic software is regulated as a medical device [118], and the government's approach to AI governance is guided by principles of safety, transparency, and accountability. However, a dedicated or unified "Law of AI" has not yet been established, resulting in a fragmented regulatory landscape that complicates questions of liability and accountability when AI-driven diagnostic recommendations are incorrect [119]. Within this context, it remains unclear who ultimately bears responsibility for errors—the physician, the system developer, or the healthcare institution.

Furthermore, the integration of XAI into clinical workflows introduces potential risks of inaccuracy, data breaches, and unintended harm. In a high-risk domain such as healthcare, diagnostic errors can have serious consequences for patients [120]. Consequently, AI-based systems should not function autonomously but rather augment the clinician's expertise and judgement. Over reliance on such technologies may contribute to clinical deskilling, shifting the physician's role from active diagnosis to passive validation of AI-generated outputs [121].

To mitigate these risks, XAI systems should be developed with a strong emphasis on human factors engineering and user-centred design, fostering what is often referred to as 'calibrated trust' between humans and machines. This approach necessitates close collaboration with practitioners to ensure that user trust is not assumed but actively measured and adjusted—increasing or decreasing trust levels as needed to maintain an appropriate balance between human oversight and algorithmic support [122].

6. Gaps and Challenges

Despite significant advances in artificial intelligence for medical imaging, several critical challenges remain that highlight the need for continued research, particularly in the context of skin cancer detection. One of the primary obstacles in multiple-lesion classification, detection, and segmentation lies in the increased complexity of these tasks when they are performed using limited and imbalanced datasets. The scarcity of high-quality annotated data restricts the generalisability of machine learning models and increases the likelihood of overfitting. This issue is illustrated by studies such as Lucieri et al. [49], where a model achieved 100% accuracy on the PH2 dataset despite being trained on only 40 images, suggesting poor generalisation beyond the training data.

In addition, existing public datasets predominantly focus on melanoma and nevi, limiting the diversity required to develop models capable of recognising a wider range of skin lesion types. Frequently used datasets, such as HAM10000, which appear in more than half of the reviewed studies, also exhibit substantial demographic bias, with a strong over representation of lighter skin tones. As a result, models trained on these datasets may demonstrate reduced diagnostic performance when applied to individuals with darker skin tones [123]. This imbalance raises significant concerns regarding the fairness, reliability, and generalisability of current explainable AI frameworks across diverse patient populations. A further limitation is the continued reliance on qualitative assessment of explanation outputs. Many studies evaluate heatmaps through visual inspection or expert judgement alone, without employing quantitative explainability metrics. While a saliency map may appear clinically plausible, visual agreement does not necessarily indicate that the explanation faithfully reflects the model's underlying decision-making process. Consequently, there is increasing recognition of the need for quantitative evaluation measures, including faithfulness, robustness, and complexity metrics, to assess whether explanations accurately represent model behaviour, remain stable under perturbations, and provide information in a manner that is both interpretable and clinically actionable. Complementary approaches such as ablation studies can further evaluate the contribution of specific features, inputs, or model components to both model predictions and generated explanations. The limited adoption of such evaluation frameworks remains a significant barrier to the rigorous assessment and comparison of XAI methods in dermatology. This challenge motivated the inclusion of robustness and fidelity as core dimensions within the GRADE-inspired evaluation framework presented in Table 7, as these criteria provide a more rigorous basis for assessing explanation quality than visual plausibility alone.

A related concern involves dataset selection and evaluation methodology. Many studies report high classification accuracy while training and testing models on the same dataset, often relying on widely used benchmarks such as HAM10000 or ISIC. Although such benchmarks are valuable for comparison, models evaluated solely within the same dataset may not generalise effectively to other clinical environments. For instance, a model trained on dermoscopic images from a particular dataset may perform considerably worse when applied to images obtained using different imaging devices, lighting conditions, or patient populations. This limitation highlights concerns regarding the real-world applicability of these systems in clinical settings.

Closely associated with this issue is the limited use of external validation. While cross-validation within a single dataset is commonly employed, comparatively few studies evaluate their models using independent external datasets. Validation on data collected from different institutions, geographical regions, or imaging devices would provide stronger evidence of model robustness and improve confidence in the potential clinical deployment of such systems.

In addition to dataset and evaluation limitations, CNN-based architectures, which underpin the majority of skin lesion classification systems identified in this review, present several methodological constraints. While CNNs are highly effective at automated feature extraction from dermoscopic images, they are prone to learning dataset-specific patterns rather than clinically meaningful representations, meaning that they may overfit to benchmark datasets. This can result in reduced generalisability when models are applied to data from different imaging devices, lighting conditions, acquisition settings, or patient populations. Second, CNNs are sensitive to artefacts such as hair, ruler markings, ink annotations, and image borders, which may be spuriously associated with malignancy labels, which can negatively impact diagnostic reliability in real-world clinical environments [124]. Third, many CNN pipelines are image-centric and do not incorporate clinically relevant metadata such as patient age, lesion location, skin type, or lesion evolution. Furthermore, CNNs provide limited intrinsic interpretability; therefore, explanations are often generated post hoc, and such explanations may not faithfully represent the model's internal decision process. These limitations underscore the importance of robust validation strategies and explainability techniques to ensure that model predictions are clinically meaningful and generalisable.

Although the field of XAI has experienced substantial growth, existing reviews frequently address explainability in healthcare in a broad sense, without focusing specifically on the unique challenges associated with skin cancer detection. Furthermore, there remains a significant need for standardised metrics to evaluate the quality, reliability, and clinical usefulness of explanations generated by XAI models [19]. These gaps indicate the importance of developing multimodal neural network architectures that integrate diverse annotated datasets with rigorous evaluation protocols, alongside explainability frameworks tailored to dermatological image analysis.

In addition, current explainability techniques present several limitations. Methods such as Grad-CAM, Grad-CAM++, LIME, and SHAP are widely used within the reviewed literature. A limitation of the reviewed literature is that many XAI outputs are evaluated visually rather than quantitatively. A heat map may appear clinically plausible while failing to faithfully represent the model's internal logic. These metrics are essential for distinguishing explanations that are merely visually persuasive from those that are causally informative. Furthermore, explanation outputs can vary depending on model architecture, parameter settings, and training conditions, raising concerns regarding their consistency and reliability. An additional challenge concerns the limited standardisation of faithfulness evaluation. Future studies should therefore incorporate quantitative XAI evaluation metrics,

including deletion and insertion tests, faithfulness correlation, localisation overlap with lesion masks, robustness under perturbation, explanation complexity or sparsity, and model-randomisation sanity checks.

Finally, practical challenges remain regarding the integration of AI-based diagnostic systems into clinical workflows. Many studies primarily focus on algorithm development and offline evaluation rather than real-world implementation. In clinical practice, however, diagnostic tools must integrate seamlessly with existing workflows, including electronic health record systems, dermatological imaging devices, and established diagnostic procedures. Addressing these practical considerations is essential for ensuring that AI-driven diagnostic systems are not only technically effective but also usable and trustworthy within real-world healthcare environment, as such, XAI is required to indicate whether the model is focusing on clinically meaningful features, such as asymmetry, border irregularity, and colour variegation, rather than artefacts such as rulers, hair, lighting, or image background. Therefore, explainability is central to converting diagnostic performance into clinically actionable support for GPs.

These findings highlight several priorities for future AI and XAI research in dermatology. First, the development of larger and more diverse datasets representing different skin tones, lesion types, and imaging modalities is essential to improve model fairness and generalisability. Second, greater emphasis should be placed on external validation using data from multiple institutions and geographical regions to establish clinical robustness. Third, standardised frameworks for evaluating explanation quality are needed to enable meaningful comparison between XAI techniques and to determine whether generated explanations align with dermatological reasoning. Finally, future systems should move beyond retrospective image analysis towards clinically integrated decision-support tools that combine high diagnostic performance with transparent, reliable, and clinically interpretable explanations.

7. Conclusions

This review highlights that contemporary research on machine learning and deep learning-based skin lesion detection has been largely driven by experimentation on a small number of benchmark dermoscopic datasets, most notably ISIC. While these datasets have enabled the development of high-performing classification models across architectures such as EfficientNet, ResNet, and Xception, the dominance of single-dataset evaluations raises significant concerns regarding dataset bias, demographic representation, and the overestimation of real-world performance. In contrast, studies incorporating multiple datasets exhibit slightly reduced accuracy than singular datasets, but demonstrate improved robustness and generalisability, underscoring the importance of moving beyond accuracy-centric evaluation toward clinically reliable validation strategies.

From an explainability perspective, the literature reveals a relatively slow and uneven integration of XAI frameworks into skin lesion classification pipelines. Early work between 2020 and 2021 relied heavily on foundational post hoc methods such as LIME, and early CAM-based approaches, which often failed to provide clinically meaningful insight and risked falling into the explainability trap. Subsequent years saw consolidation around Grad-CAM, SHAP, and LIME, reflecting a degree of methodological maturity but also a growing dependence on a narrow set of tools not originally designed for medical imaging. From 2023 onwards, a gradual diversification emerged, with the introduction of refined and complementary methods such as Grad-CAM++, Score-CAM, occlusion-based approaches, dimensionality-reduction techniques, and prototype-based networks signalling an expanding but still fragmented explainability landscape. While saliency-based methods have improved transparency by revealing where models focus their attention,

true clinical adoption may depend on explainability techniques that communicate why a prediction was made in terms that are meaningful to clinicians. A central conclusion of this review is that clinically useful XAI must move beyond pixel-level saliency. For GP-facing dermatology tools, explanations should be expressed in terms of clinically meaningful descriptors such as asymmetry, border irregularity, colour variation, lesion diameter, lesion evolution, pigment network, streaks, and blue-white veil. Until XAI systems communicate in the language of dermatological reasoning rather than only highlighting pixel clusters, their ability to support safe and calibrated clinical adoption will remain limited.

Despite these developments, significant limitations persist. The majority of systems remain image-centric, with limited incorporation of clinical metadata or multimodal inputs, restricting both diagnostic scope and explanatory depth. Furthermore, XAI methods are often evaluated qualitatively, with insufficient assessment of their stability, fidelity, or alignment with dermatological reasoning. The lack of standardised evaluation protocols further complicates comparisons across studies and hinders translation into clinical practice.

Importantly, only a small subset of studies meaningfully involve clinicians or patients in the design or evaluation of explainable systems. Where such engagement exists, findings consistently indicate that explanation quality directly influences trust, diagnostic confidence, and willingness to adopt AI-assisted tools. However, systematic frameworks for capturing and operationalising this feedback remain underdeveloped.

Taken together, these findings emphasise a necessary shift in the field from optimising performance on isolated datasets toward developing explainability-first, multimodal, and clinically grounded AI systems evaluated across diverse populations and data sources. Future research must prioritise cross-dataset validation, richer contextual explanations, and rigorous assessment of explainability techniques in real-world clinical settings. In response to these challenges, our forthcoming work aims to integrate dermoscopic imagery with clinical metadata within an explainability-driven model architecture, with the goal of delivering transparent, dermatologically meaningful insights that better support clinician decision-making and foster calibrated trust in AI-assisted diagnostics.

Author Contributions: Conceptualization, J.K. and J.W.; methodology, J.K., J.W. and R.P.; validation, J.K. and R.P.; investigation and formal analysis, J.W. and J.K.; resources, J.K. and R.P.; data curation, J.W.; writing—original draft preparation, J.W. and J.K.; writing—review and editing, J.K. and R.P.; visualization, J.W. and J.K.; supervision and project administration, J.K. and R.P.; funding acquisition, J.K. and R.P. All authors have read and agreed to the published version of the manuscript.

Funding: This project is partly supported by the University of South Wales (USW), Faculty of Computing, Engineering and Science (Ref: 30002739), and Cardiff & Vale University Health Board, Rehabilitation Engineering Unit (Ref: 102677.1122).

Institutional Review Board Statement: Not applicable.

Informed Consent Statement: Not applicable.

Data Availability Statement: No new data were created or analyzed in this study. Data sharing is not applicable to this article.

Acknowledgments: The authors would like to thank the University of South Wales and Cardiff and Vale University Health Board, REU, for supporting the project. During the preparation of this manuscript, the author(s) used ChatGPT (OpenAI) (version 5.3) solely for language editing and typographical/grammar checking. The authors reviewed and revised the text and take full responsibility for the content of this publication.

Conflicts of Interest: The authors declare no conflicts of interest.

References

1. NHS England. *Referral Optimisation for People with Skin Conditions*; NHS England: London, UK, 2022.
2. Fee, J.A.; McGrady, F.P.; Hart, N.D. Dermoscopy use in primary care: A qualitative study with general practitioners. *BMC Prim. Care* **2022**, *23*, 47. [CrossRef]
3. NHS England Digital. *Urgent Suspected Cancer Referrals: Referral, Conversion and Detection Rates, April 2021 to March 2023*; NHS England: London, UK, 2024.
4. Kownacki, S. Skin diseases in primary care: What should GPs be doing? *Br. J. Gen. Pract.* **2014**, *64*, 380–381. [CrossRef] [PubMed]
5. Harkemanne, E.; Duyver, C.; Leconte, S.; Bugli, C.; Thomas, L.; Baeck, M.; Tromme, I. Melanoma Diagnostic Practices of French-Speaking Belgian General Practitioners and the Prospective Study of Their Pigmented Skin Lesion Diagnostic Accuracy and Management. *J. Cancer Educ.* **2020**, *36*, 1316–1324. [CrossRef] [PubMed]
6. Shetty, B.; Fernandes, R.; Rodrigues, A.P.; Chengoden, R.; Bhattacharya, S.; Lakshmana, K. Skin lesion classification of dermoscopic images using machine learning and convolutional neural network. *Sci. Rep.* **2022**, *12*, 18134. [CrossRef] [PubMed]
7. Brinker, T.J.; Hekler, A.; Enk, A.H.; Klode, J.; Hauschild, A.; Berking, C.; Schilling, B.; Haferkamp, S.; Schadendorf, D.; Holland-Letz, T.; et al. Deep learning outperformed 136 of 157 dermatologists in a head-to-head dermoscopic melanoma image classification task. *Eur. J. Cancer* **2019**, *113*, 47–54. [CrossRef] [PubMed]
8. Haenssle, H.A.; Fink, C.; Schneiderbauer, R.; Toberer, F.; Buhl, T.; Blum, A.; Kalloo, A.; Hassen, A.B.H.; Thomas, L.; Enk, A.; et al. Man against machine: Diagnostic performance of a deep learning convolutional neural network for dermoscopic melanoma recognition in comparison to 58 dermatologists. *Ann. Oncol.* **2018**, *29*, 1836–1842. [CrossRef] [PubMed]
9. Tschandl, P.; Rinner, C.; Apalla, Z.; Argenziano, G.; Codella, N.; Halpern, A.; Janda, M.; Lallas, A.; Longo, C.; Malvehy, J.; et al. Human–computer collaboration for skin cancer recognition. *Nat. Med.* **2020**, *26*, 1229–1234. [CrossRef] [PubMed]
10. Esteva, A.; Kuprel, B.; Novoa, R.A.; Ko, J.; Swetter, S.M.; Blau, H.M.; Thrun, S. Dermatologist-level Classification of Skin Cancer with Deep Neural Networks. *Nature* **2017**, *542*, 115–118. [CrossRef] [PubMed]
11. Metta, C.; Beretta, A.; Guidotti, R.; Yin, Y.; Gallinari, P.; Rinzivillo, S.; Giannotti, F. Explainable Deep Image Classifiers for Skin Lesion Diagnosis. *arXiv* **2021**, arXiv:2111.11863.
12. Yu, Y.; Jia, H.; Zhang, L.; Xu, S.; Zhu, X.; Wang, J.; Wang, F.; Han, L.; Jiang, H.; Zhou, Q.; et al. Deep Multi-Modal Skin-Imaging-Based Information-Switching Network for Skin Lesion Recognition. *Bioengineering* **2025**, *12*, 282. [CrossRef] [PubMed]
13. Dillshad, V.; Khan, M.A.; Nazir, M.; Ahmad, J.; AlHammadi, D.A.; Houda, T.; Cho, H.C.; Chang, B. A Multi-Layers Information Fused Deep Architecture for Skin Cancer Classification in Smart Healthcare. *Comput. Mater. Contin.* **2025**, *83*, 5299–5321. [CrossRef]
14. Nemade, G. How to Explain and Interpret Machine Learning Models with Explainable AI. Medium, 2025. Available online: <https://medium.com/@nemagan/how-to-explain-and-interpret-machine-learning-models-with-explainable-ai-904e11745066> (accessed on 24 March 2026).
15. Singh, Y.; Hathaway, Q.A.; Keishing, V.; Salehi, S.; Wei, Y.; Horvat, N.; Vera-Garcia, D.V.; Choudhary, A.; Kh, A.M.; Quaia, E.; et al. Beyond Post hoc Explanations: A Comprehensive Framework for Accountable AI in Medical Imaging Through Transparency, Interpretability, and Explainability. *Bioengineering* **2025**, *12*, 879–879. [CrossRef] [PubMed]
16. Ali, S.; Abuhmed, T.; El-Sappagh, S.; Muhammad, K.; Alonso-Moral, J.M.; Confalonieri, R.; Guidotti, R.; Ser, J.D.; Díaz-Rodríguez, N.; Herrera, F. Explainable Artificial Intelligence (XAI): What we know and what is left to attain Trustworthy Artificial Intelligence. *Inf. Fusion* **2023**, *99*, 101805. [CrossRef]
17. Kreps, S.E.; George, J.; Lushenko, P.; Rao, A.S. Exploring the Artificial Intelligence “Trust paradox”: Evidence from a Survey Experiment in the United States. *PLoS ONE* **2023**, *18*, e0288109. [CrossRef] [PubMed]
18. Turner, A.; Kaushik, M.; Huang, M.-T.; Varanasi, S. *Calibrating Trust in AI-Assisted Decision Making*; MIMS Capstone Project Report; School of Information, University of California, Berkeley: Berkeley, CA, USA, 2020. Available online: <https://www.ischool.berkeley.edu/projects/2020/calibrating-trust-ai-assisted-decision-making> (accessed on 24 March 2026).
19. Gupta, J.; Seeja, K.R. A Comparative Study and Systematic Analysis of XAI Models and their Applications in Healthcare. *Arch. Comput. Methods Eng.* **2024**, *31*, 3977–4002. [CrossRef]
20. Farrugia, P.; Petrisor, B.A.; Farrokhyar, F.; Bhandari, M. Research questions, hypotheses and objectives. *Can. J. Surg.* **2010**, *53*, 278. [PubMed]
21. PRISMA. *PRISMA Statement*; PRISMA: Easley, SC, USA, 2020.
22. Haddaway, N.R.; Page, M.J.; Pritchard, C.C.; McGuinness, L.A. PRISMA2020: An R Package and Shiny App for Producing PRISMA 2020-compliant Flow diagrams, with Interactivity for Optimised Digital Transparency and Open Synthesis. *Campbell Syst. Rev.* **2022**, *18*, e1230. [CrossRef]
23. American Academy of Dermatology Association. Types of Skin Cancer. 2024. Available online: <https://www.aad.org/public/diseases/skin-cancer/types/common> (accessed on 24 March 2026).
24. Davis, L.E.; Shalin, S.C.; Tackett, A.J. Current state of melanoma diagnosis and treatment. *Cancer Biol. Ther.* **2019**, *20*, 1366–1379. [CrossRef] [PubMed]

25. Walter, F.M.; Prevost, A.T.; Vasconcelos, J.; Hall, P.N.; Burrows, N.P.; Morris, H.C.; Kinmonth, A.L.; Emery, J.D. Using the 7-point checklist as a diagnostic aid for pigmented skin lesions in general practice: A diagnostic validation study. *Br. J. Gen. Pract.* **2013**, *63*, e345–e353. [CrossRef] [PubMed]
26. Cancer Research UK. Melanoma Skin Cancer Statistics. Cancer Research UK. 2024. Available online: <https://www.cancerresearchuk.org/health-professional/cancer-statistics/statistics-by-cancer-type/melanoma-skin-cancer> (accessed on 24 March 2026).
27. Medicolegal Partners. The Medicolegal Challenges of Skin Cancer Management. Medicolegal Partners. 2024. Available online: <https://www.medicolegal-partners.com/the-medicolegal-challenges-of-skin-cancer-management/> (accessed on 24 March 2026).
28. Spyridonos, P.; Gaitanis, G.; Likas, A.; Bassukas, I. Characterizing Malignant Melanoma Clinically Resembling Seborrheic Keratosis Using Deep Knowledge Transfer. *Cancers* **2021**, *13*, 6300. [CrossRef] [PubMed]
29. Jensen, J.D.; Elewski, B.E. The ABCDEF Rule: Combining the “ABCDE Rule” and the “Ugly Duckling Sign” in an Effort to Improve Patient Self-Screening Examinations. *J. Clin. Aesthetic Dermatol.* **2015**, *8*, 15.
30. Bell, K.J.; Nijsten, T. Melanoma overdiagnosis: Why it matters and what can be done about it. *Br. J. Dermatol.* **2022**, *187*, 459. [CrossRef] [PubMed]
31. Nigar, N.; Umar, M.; Shahzad, M.K.; Islam, S.; Abalo, D. A Deep Learning Approach Based on Explainable Artificial Intelligence for Skin Lesion Classification. *IEEE Access* **2022**, *10*, 113715–113725. [CrossRef]
32. Efat, A.H.; Mahedy, M.; Uddin, M.P.; Mamun, M.A. A Multi-level ensemble approach for skin lesion classification using Customized Transfer Learning with Triple Attention. *PLoS ONE* **2024**, *19*, e0309430. [CrossRef] [PubMed]
33. Efat, A.H. Chi2 weighted ensemble: A multi-layer ensemble approach for skin lesion classification using a novel framework—Optimized RegNet synergy with Attention-Triplet. *PLoS ONE* **2025**, *20*, e0321803. [CrossRef] [PubMed]
34. Ajmal, M.; Khan, M.A.; Akram, T.; Alqahtani, A.; Alhaisoni, M.; Armghan, A.; Althubiti, S.A.; Alenezi, F. BF2SkNet: Best deep learning features fusion-assisted framework for multiclass skin lesion classification. *Neural Comput. Appl.* **2022**, *35*, 22115–22131. [CrossRef]
35. AbuAlkebash, H.; Saleh, R.A.; Ertunc, H.M. Automated explainable deep learning framework for multiclass skin cancer detection and classification using hybrid YOLOv8 and vision transformer (ViT). *Biomed. Signal Process. Control* **2025**, *108*, 107934. [CrossRef]
36. Meena, J.; Hasija, Y. Application of explainable artificial intelligence in the identification of Squamous Cell Carcinoma biomarkers. *Comput. Biol. Med.* **2022**, *146*, 105505. [CrossRef] [PubMed]
37. Mridha, K.; Uddin, M.M.; Shin, J.; Khadka, S.; Mridha, M.F. An Interpretable Skin Cancer Classification Using Optimized Convolutional Neural Network for a Smart Healthcare System. *IEEE Access* **2023**, *11*, 41003–41018. [CrossRef]
38. Jha, A.; Ananthanarayana, V.S. An Interpretable Deep Learning Model for Skin Lesion Classification. In *Communications in Computer and Information Science*; Springer Nature: Cham, Switzerland, 2023; pp. 543–553. [CrossRef]
39. Biswas, S.; Saha, S.; Uddin, M.S.; Mostafiz, R. An explainable and federated deep learning framework for skin cancer diagnosis. *PLoS ONE* **2025**, *20*, e0324393. [CrossRef] [PubMed]
40. Öznacar, T.; Varol Kayapunar, N. Advanced skin cancer prediction with medical image data using MobileNetV2 deep learning and optimized techniques. *Sci. Rep.* **2025**, *15*, 28962. [CrossRef] [PubMed]
41. Mahmud, M.Z.; Reza, M.S.; Alve, S.R.; Islam, S.; Fahmid, N. Advance Transfer Learning Approach for Identification of Multiclass Skin Disease with LIME Explainable AI Technique. In *2024 27th International Conference on Computer and Information Technology (ICCIT)*; IEEE: Piscataway, NJ, USA, 2024; pp. 109–114. [CrossRef]
42. Thwin, S.M.; Park, H.S.; Seo, S.H. A Trustworthy Framework for Skin Cancer Detection Using a CNN with a Modified Attention Mechanism. *Appl. Sci.* **2025**, *15*, 1067. [CrossRef]
43. Ahmad, N.; Shah, J.H.; Khan, M.A.; Baili, J.; Ansari, G.J.; Tariq, U.; Kim, Y.J.; Cha, J.H. A novel framework of multiclass skin lesion recognition from dermoscopic images using deep learning and explainable AI. *Front. Oncol.* **2023**, *13*, 1151257. [CrossRef] [PubMed]
44. Sasithradevi, A.; Kanimozhi, S.; Sasidhar, P.; Kumar Pulipati, P.; Sruthi, E.; Prakash, P. EffiCAT: A synergistic approach to skin disease classification through multi-dataset fusion and attention mechanisms. *Biomed. Signal Process. Control* **2024**, *100*, 107141. [CrossRef]
45. Hachi, S.; Sabri, A. Development of an Explainable AI (XAI) system for the interpretation of ViT in the diagnosis of medical images. In *2024 3rd International Conference on Embedded Systems and Artificial Intelligence (ESAI)*; IEEE: Piscataway, NJ, USA, 2024; pp. 1–8. [CrossRef]
46. Chand, H.; Kumar, D.; Bali, K.K. Enhancing Interpretability of Skin Lesion Classification Using Grad-CAM and Weighted Grad-CAM. In *2024 International Conference on Sustainable Technology and Engineering (i-COSTE)*; IEEE: Piscataway, NJ, USA, 2024; pp. 1–7. [CrossRef]
47. Mohan, J.; Sivasubramanian, A.; V., S.; Ravi, V. Enhancing skin disease classification leveraging transformer-based deep learning architectures and explainable AI. *Comput. Biol. Med.* **2025**, *190*, 110007. [CrossRef] [PubMed]
48. Choi, J.Y.; Song, M.J.; Shin, Y.J. Enhancing Skin Lesion Classification Performance with the ABC Ensemble Model. *Appl. Sci.* **2024**, *14*, 10294. [CrossRef]

49. Lucieri, A.; Bajwa, M.N.; Braun, S.A.; Malik, M.I.; Dengel, A.; Ahmed, S. ExAID: A multimodal explanation framework for computer-aided diagnosis of skin lesions. *Comput. Methods Programs Biomed.* **2022**, *215*, 106620. [\[CrossRef\]](#) [\[PubMed\]](#)
50. Shah, S.; Shah, S.; Khaled, R.; Buccoliero, A.; Shah, S.; Di Terlizzi, A.; Di Benedetto, G.; Deriu, M. Explainable AI-Based Skin Cancer Detection Using CNN, Particle Swarm Optimization and Machine Learning. *J. Imaging* **2024**, *10*, 332. [\[CrossRef\]](#) [\[PubMed\]](#)
51. Abdullah, M.; Afrin, S.; Mridha, M.F.; Alfarhood, S.; Che, D.; Safran, M. Explainable deep learning approaches for high precision early melanoma detection using dermoscopic images. *Sci. Rep.* **2025**, *15*, 24533. [\[CrossRef\]](#) [\[PubMed\]](#)
52. Abdulredah, A.A.; Fadhel, M.A.; Alzubaidi, L.; Duan, Y.; Kherallah, M.; Charfi, F. Towards unbiased skin cancer classification using deep feature fusion. *BMC Med. Inform. Decis. Mak.* **2025**, *25*, 48. [\[CrossRef\]](#) [\[PubMed\]](#)
53. Arshad, M.; Khan, M.A.; Almujaally, N.A.; Alasiry, A.; Marzougui, M.; Nam, Y. Multiclass skin lesion classification and localization from dermoscopic images using a novel network-level fused deep architecture and explainable artificial intelligence. *BMC Med. Inform. Decis. Mak.* **2025**, *25*, 215. [\[CrossRef\]](#) [\[PubMed\]](#)
54. Alrabai, A.; Echtioui, A.; Kallel, F. Exploring Pre-Trained Models for Skin Cancer Classification. *Appl. Syst. Innov.* **2025**, *8*, 35. [\[CrossRef\]](#)
55. Cino, L.; Distanto, C.; Martella, A.; Mazzeo, P.L. Skin Lesion Classification Through Test Time Augmentation and Explainable Artificial Intelligence. *J. Imaging* **2025**, *11*, 15. [\[CrossRef\]](#) [\[PubMed\]](#)
56. Alqarni, A.N.; Sheikh, A. SkinDiseaseXAI: XAI-Driven Neural Networks for Skin Disease Detection. *Int. J. Adv. Comput. Sci. Appl.* **2025**, *16*, 96–104. [\[CrossRef\]](#)
57. Roy, A.L.; Maruf, A.A.; Sony, R.I.; Roy, A.P.; Noor, N.; Aung, Z. Fostering trust and interpretability in skin cancer classification: A hybrid framework of deep learning and machine learning with explainable AI. *J. Ambient Intell. Humaniz. Comput.* **2025**, *16*, 879–894. [\[CrossRef\]](#)
58. Oyediji, M.O.; Okafor, E.; Samma, H.; Alfarraj, M. Interpretable Deep Learning for Classifying Skin Lesions. *Int. J. Intell. Syst.* **2025**, *2025*, 2751767. [\[CrossRef\]](#)
59. Ali, A.; Shahbaz, H.; Damaševičius, R. xCViT: Improved Vision Transformer Network with Fusion of CNN and Xception for Skin Disease Recognition with Explainable AI. *Comput. Mater. Contin.* **2025**, *83*, 1367–1398. [\[CrossRef\]](#)
60. Munjal, G.; Bhardwaj, P.; Bhargava, V.; Singh, S.; Nagpal, N. SkinSage XAI: An explainable deep learning solution for skin lesion diagnosis. *Health Care Sci.* **2024**, *3*, 438–455. [\[CrossRef\]](#) [\[PubMed\]](#)
61. Attallah, O. Skin-CAD: Explainable deep learning classification of skin cancer from dermoscopic images by feature selection of dual high-level CNNs features and transfer learning. *Comput. Biol. Med.* **2024**, *178*, 108798. [\[CrossRef\]](#) [\[PubMed\]](#)
62. Gamage, L.; Isuranga, U.; Meedeniya, D.; De Silva, S.; Yogarajah, P. Melanoma Skin Cancer Identification with Explainability Utilizing Mask Guided Technique. *Electronics* **2024**, *13*, 680. [\[CrossRef\]](#)
63. Joseph, S.; Joshi, H.; Singh, S.; Mayekar, O.; Wagh, M. Hybrid VGG-SVM Framework for Melanoma Detection: Integrating GAN-Augmented Data and LIME Interpretability. In *2024 2nd DMIHER International Conference on Artificial Intelligence in Healthcare, Education and Industry (IDICAIEI)*; IEEE: Piscataway, NJ, USA, 2024; pp. 1–8. [\[CrossRef\]](#)
64. Hasib, M.M.; Faruk, M.F.; Mahedy, M.; Mamun, M.A.; Srizon, A.Y.; Efat, A.H.; Hossain, M.R.; Khansur, M.N. Improved Skin Lesion Detection with Double Layer Concatenated DenseNet Using Transfer Learning and Attention Modules. In *2024 IEEE International Conference on Power, Electrical, Electronics and Industrial Applications (PEEIACON)*; IEEE: Piscataway, NJ, USA, 2024; pp. 1–6. [\[CrossRef\]](#)
65. Sah, N.K.; Reddy, S.; Ullas, K.; Singh, T.; Mandal, A.; Chatterji, S. Interpretable Deep Learning for Skin Cancer Detection: Exploring LIME and SHAP. In *2022 13th International Conference on Computing Communication and Networking Technologies (ICCCNT)*; IEEE: Piscataway, NJ, USA, 2024; pp. 1–7. [\[CrossRef\]](#)
66. Dagnaw, G.; Mouhtadi, E.; Mustapha, M. Skin cancer classification using vision transformers and explainable artificial intelligence. *J. Med. Artif. Intell.* **2024**, *7*, 1. [\[CrossRef\]](#)
67. Ulla, S.; Yousuf, M.A. MSCC: Multi-Class Skin Cancer Classification and Interpretable Deep Learning Systems. In *2024 6th International Conference on Electrical Engineering and Information & Communication Technology (ICEEICT)*; IEEE: Piscataway, NJ, USA, 2024. [\[CrossRef\]](#)
68. Malhotra, N.; Kaur, P. Exploring Explainable Melanoma Classification: Leveraging Pre-trained Deep Learning Model on MED-NODE Dataset. In *2024 11th International Conference on Computing for Sustainable Global Development (INDIACom)*; IEEE: Piscataway, NJ, USA, 2024. [\[CrossRef\]](#)
69. Effendie, M.; Aguiar-Pulido, V. Multi-Modal AI Approach for Multi-Class Skin Disease Classification. In *2024 International Conference on Machine Learning and Applications (ICMLA)*; IEEE: Piscataway, NJ, USA, 2024; pp. 1588–1592. [\[CrossRef\]](#)
70. Khater, T.; Ansari, S.; Mahmoud, S.A.; Hussain, A.J.; Tawfik, H. Skin Cancer Classification Using Explainable Artificial Intelligence on Pre-extracted Image Features. *Intell. Syst. Appl.* **2023**, *20*, 200275. [\[CrossRef\]](#)
71. Gamage, L.; Isuranga, U.; Silva, S.D.; Meedeniya, D. Melanoma Skin Cancer Classification with Explainability. In *2023 3rd International Conference on Advanced Research in Computing (ICARC)*; IEEE: Piscataway, NJ, USA, 2023. [\[CrossRef\]](#)

72. Athina, F.H.; Sara, S.A.; Sarwar, Q.S.; Tabassum, N.; Era, M.T.; Ashraf, F.B.; Hossain, M.I. Multi-classification Network for Detecting Skin Diseases using Deep Learning and XAI. In *2021 International Conference on Innovation and Intelligence for Informatics, Computing, and Technologies (3ICT)*; IEEE: Piscataway, NJ, USA, 2022; pp. 648–655. [\[CrossRef\]](#)
73. Pintelas, E.; Liaskos, M.; Livieris, I.E.; Kotsiantis, S.; Pintelas, P. A novel explainable image classification framework: Case study on skin cancer and plant disease prediction. *Neural Comput. Appl.* **2021**, *33*, 15171–15189. [\[CrossRef\]](#)
74. Ganguly, B.; Dey, D.; Munshi, S. Image Visibility Filter-Based Interpretable Deep Learning Framework for Skin Lesion Diagnosis. *IEEE Trans. Ind. Inform.* **2021**, *18*, 5138–5147. [\[CrossRef\]](#)
75. Jaworek-Korjakowska, J.; Brodzicki, A.; Cassidy, B.; Kendrick, C.; Yap, M.H. Interpretability of a Deep Learning Based Approach for the Classification of Skin Lesions into Main Anatomic Body Sites. *Cancers* **2021**, *13*, 6048. [\[CrossRef\]](#) [\[PubMed\]](#)
76. Wang, S.; Yin, Y.; Wang, D.; Wang, Y.; Jin, Y. Interpretability-Based Multimodal Convolutional Neural Networks for Skin Lesion Diagnosis. *IEEE Trans. Cybern.* **2022**, *52*, 12623–12637. [\[CrossRef\]](#) [\[PubMed\]](#)
77. Lucieri, A.; Bajwa, M.N.; Alexander Braun, S.; Malik, M.I.; Dengel, A.; Ahmed, S. On Interpretability of Deep Learning based Skin Lesion Classifiers using Concept Activation Vectors. In *2020 International Joint Conference on Neural Networks (IJCNN)*; IEEE: Piscataway, NJ, USA, 2020. [\[CrossRef\]](#)
78. Nunnari, F.; Kadir, M.A.; Sonntag, D. On the Overlap Between Grad-CAM Saliency Maps and Explainable Visual Features in Skin Cancer Images. *Lect. Notes Comput. Sci.* **2021**, *12844*, 241–253. [\[CrossRef\]](#)
79. Stieler, F.; Rabe, F.; Bauer, B. Towards Domain-Specific Explainable AI: Model Interpretation of a Skin Image Classifier using a Human Approach. In *2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)*; IEEE: Piscataway, NJ, USA, 2021. [\[CrossRef\]](#)
80. Xiang, A.; Wang, F. Towards Interpretable Skin Lesion Classification with Deep Learning Models. *AMIA Annu. Symp. Proc.* **2020**, *2019*, 1246. [\[PubMed\]](#)
81. Copeland, B.J. Artificial Intelligence. Encyclopaedia Britannica. 2024. Available online: <https://www.britannica.com/technology/artificial-intelligence> (accessed on 24 March 2026).
82. Betz, S. 7 Types of Artificial Intelligence. Built In. 2022. Available online: <https://builtin.com/artificial-intelligence/types-of-artificial-intelligence> (accessed on 24 March 2026).
83. Janiesch, C.; Zschech, P.; Heinrich, K. Machine learning and deep learning. *Electron. Mark.* **2021**, *31*, 685–695. [\[CrossRef\]](#)
84. Jiang, H.; Diao, Z.; Shi, T.; Zhou, Y.; Wang, F.; Hu, W.; Zhu, X.; Luo, S.; Tong, G.; Yao, Y.D. A review of deep learning-based multiple-lesion recognition from medical images: Classification, detection and segmentation. *Comput. Biol. Med.* **2023**, *157*, 106726. [\[CrossRef\]](#) [\[PubMed\]](#)
85. Kosinski, M. What Is Black Box Artificial Intelligence (AI)? 2024. Available online: <https://www.ibm.com/think/topics/black-box-ai> (accessed on 24 March 2026).
86. Linardatos, P.; Papastefanopoulos, V.; Kotsiantis, S. Explainable AI: A Review of Machine Learning Interpretability Methods. *Entropy* **2020**, *23*, 18. [\[CrossRef\]](#) [\[PubMed\]](#)
87. Madathil, A.P.; Luo, X.; Liu, Q.; Walker, C.; Madarkar, R.; Cai, Y.; Liu, Z.; Chang, W.; Qin, Y. Intrinsic and post-hoc XAI approaches for fingerprint identification and response prediction in smart manufacturing processes. *J. Intell. Manuf.* **2024**, *35*, 4159–4180. [\[CrossRef\]](#)
88. Retzlaff, C.O.; Angerschmid, A.; Saranti, A.; Schneeberger, D.; Röttger, R.; Müller, H.; Holzinger, A. Post-hoc vs ante-hoc explanations: XAI design guidelines for data scientists. *Cogn. Syst. Res.* **2024**, *86*, 101243. [\[CrossRef\]](#)
89. Santiago, C.; Correia, M.; Verdelho, M.; Bissoto, A.; Barata, C. *Global and Local Explanations for Skin Cancer Diagnosis Using Prototypes*; Springer Nature: Cham, Switzerland, 2023.
90. Selvaraju, R.R.; Cogswell, M.; Das, A.; Vedantam, R.; Parikh, D.; Batra, D. Grad-CAM: Visual Explanations from Deep Networks via Gradient-based Localization. *Int. J. Comput. Vis.* **2020**, *128*, 336–359. [\[CrossRef\]](#)
91. Goel, A. Model Explainability Using SHAP (SHapley Additive exPlanations) and LIME (Local Interpretable Model-Agnostic Explanations). Medium, 2023. Available online: <https://medium.com/@anshulgoel991/model-explainability-using-shap-shapley-additive-explanations-and-lime-local-interpretable-cb4f5594fc1a> (accessed on 24 March 2026).
92. Roberts, A. Explainability In Machine Learning: Top Techniques. Arize AI. 2024. Available online: <https://arize.com/blog-course/explainability-techniques-shap/> (accessed on 24 March 2026).
93. Salih, A.M.; Estabragh, Z.R.; Galazzo, I.B.; Radeva, P.; Petersen, S.E.; Lekadir, K.; Menegaz, G. A Perspective on Explainable Artificial Intelligence Methods: SHAP and LIME. *Adv. Intell. Syst.* **2024**, *7*, 2400304. [\[CrossRef\]](#)
94. 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. *ACM Comput. Surv.* **2023**, *55*, 295. [\[CrossRef\]](#)
95. Kadir, A.; Mosavi, A.; Sonntag, D. Evaluation Metrics for XAI: A Review, Taxonomy, and Practical Applications. In *2023 IEEE 27th International Conference on Intelligent Engineering Systems (INES)*; IEEE: Piscataway, NJ, USA, 2023.

96. Yamauchi, T. Spatial Sensitive Grad-CAM++: Improved Visual Explanation for Object Detectors via Weighted Combination of Gradient Map. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) Workshops, Nashville, TN, USA, 20–25 June 2021.
97. Arenas, M.; Barceló, P.; Bertossi, L.; Monet, M. On the Complexity of SHAP-Score-Based Explanations: Tractability via Knowledge Compilation and Non-Approximability Results. *J. Mach. Learn. Res.* **2023**, *24*, 1–58.
98. Balshem, H.; Helfand, M.; Schünemann, H.J.; Oxman, A.D.; Kunz, R.; Brozek, J.; Vist, G.E.; Falck-Ytter, Y.; Meerpohl, J.; Norris, S. GRADE Guidelines: 3. Rating the Quality of Evidence. *J. Clin. Epidemiol.* **2011**, *64*, 401–406. [\[CrossRef\]](#) [\[PubMed\]](#)
99. Kantz, B.; Staudinger, C.; Feilmayr, C.; Wachlmayr, J.; Haberl, A.; Schuster, S.; Pernkopf, F. Robustness of Explainable Artificial Intelligence in Industrial Process Modelling. *arXiv* **2024**, arXiv:2407.09127. [\[CrossRef\]](#)
100. Jin, W.; Li, X.; Hamarneh, G. Why is Plausibility Surprisingly Problematic as an XAI Criterion? *arXiv* **2023**, arXiv:2303.17707. [\[CrossRef\]](#)
101. Miró-Nicolau, M.; Jaume-i Capó, A.; Moyà-Alcover, G. A Comprehensive Study on Fidelity Metrics for XAI. *Inf. Process. Manag.* **2024**, *62*, 103900. [\[CrossRef\]](#)
102. Haque, A.B.; Islam, A.K.M.N.; Mikalef, P. Explainable Artificial Intelligence (XAI) from a User Perspective: A Synthesis of Prior Literature and Problematising Avenues for Future Research. *Technol. Forecast. Soc. Change* **2023**, *186*, 122120. [\[CrossRef\]](#)
103. Matas, I. AI-Driven Skin Cancer Diagnosis: Grad-CAM and Expert Annotations for Enhanced Interpretability. *arXiv* **2024**, arXiv:2407.00104.
104. Giavina-Bianchi, M.; Vitor, W.G.; Fornasiero, V.; Okita, A.L.; Sousa, R.M.; Machado, B. Explainability Agreement between Dermatologists and Five Visual Explanations Techniques in Deep Neural Networks for Melanoma AI Classification. *Front. Med.* **2023**, *10*, 1. [\[CrossRef\]](#) [\[PubMed\]](#)
105. Aysel, H.I.; Cai, X.; Prugel-Bennett, A. Explainable Artificial Intelligence: Advancements and Limitations. *Appl. Sci.* **2025**, *15*, 7261. [\[CrossRef\]](#)
106. Slack, D.; Hilgard, S.; Jia, E.; Singh, S.; Lakkaraju, H. Fooling LIME and SHAP: Adversarial Attacks on Post Hoc Explanation Methods. In Proceedings of the AAAI/ACM Conference on AI, Ethics, and Society, New York, NY, USA, 7–8 February 2020; pp. 180–186. [\[CrossRef\]](#)
107. Correia, M.; Bissoto, A.; Santiago, C.; Barata, C. XAI for Skin Cancer Detection with Prototypes and Non-Expert Supervision. *arXiv* **2024**, arXiv:2402.01410. [\[CrossRef\]](#)
108. Mukherjee, S.; Chandrakar, K.; Chowdhury, S.; Senniappan, V.; Roy, B.; Panigrahi, A.; Pati, A.; Sahu, B.; Kant, S. Explainable Deep Learning for Skin Cancer Detection Using Swish-Activated Convolutional Networks. *Discov. Oncol.* **2026**, *17*, 256. [\[CrossRef\]](#) [\[PubMed\]](#)
109. Schwinghammer, M.C.; Schmalwasser, L.; Chamarthi, S.; Shardt, Y.A.W. Generalizability of Concept Knowledge in Machine Learning Using TCAV Scores: A Case Study Using Different Skin-Lesion Datasets. *IFAC-PapersOnLine* **2025**, *59*, 397–402. [\[CrossRef\]](#)
110. Patricio, C.; Teixeira, L.F.; Neves, J.C. Towards Concept-Based Interpretability of Skin Lesion Diagnosis Using Vision-Language Models. In *2024 IEEE International Symposium on Biomedical Imaging (ISBI)*; IEEE: Piscataway, NJ, USA, 2024; pp. 1–5. [\[CrossRef\]](#)
111. Yu, Z.; Nguyen, T.D.; Ju, L.; Gal, Y.; Sashindranath, M.; Bonnington, P.; Zhang, L.; Mar, V.; Ge, Z. Hierarchical Skin Lesion Image Classification with Prototypical Decision Tree. *npj Digit. Med.* **2025**, *8*, 26. [\[CrossRef\]](#) [\[PubMed\]](#)
112. Akinrinade, O.; Du, C. Skin Cancer Detection Using Deep Machine Learning Techniques. *Intell.-Based Med.* **2024**, *11*, 100191. [\[CrossRef\]](#)
113. Ashtagi, R.; Kotecha, K.; Padthe, A.; Shinde, S.; Chinchmalatpure, S.; Mane, D. Design of an AI Layer for Real-Time Skin Cancer Diagnosis. *Rev. D'Intell. Artif.* **2024**, *38*, 377–385. [\[CrossRef\]](#)
114. Farea, E.; Saleh, R.A.A.; AbuAlkebash, H.; Farea, A.A.R.; Al-Antari, M.A. A Hybrid Deep Learning Skin Cancer Prediction Framework. *Eng. Sci. Technol. Int. J.* **2024**, *57*, 101818. [\[CrossRef\]](#)
115. Jutte, L.; Gonzalez-Villa, S.; Quintana, J.; Steven, M.; Garcia, R.; Roth, B. Integrating Generative AI with ABCDE Rule Analysis for Enhanced Skin Cancer Diagnosis, Dermatologist Training and Patient Education. *Front. Med.* **2024**, *11*, 1445318. [\[CrossRef\]](#) [\[PubMed\]](#)
116. Metta, C.; Beretta, A.; Guidotti, R.; Yin, Y.; Gallinari, P.; Rinzivillo, S.; Giannotti, F. Improving Trust and Confidence in Medical Skin Lesion Diagnosis through Explainable Deep Learning. *Int. J. Data Sci. Anal.* **2023**, *20*, 183–195. [\[CrossRef\]](#)
117. Balasubramaniam, N.; Kauppinen, M.; Rannisto, A.; Hiekkänen, K.; Kujala, S. Transparency and Explainability of AI Systems: From Ethical Guidelines to Requirements. *Inf. Softw. Technol.* **2023**, *159*, 107197. [\[CrossRef\]](#)
118. Medicines and Healthcare Products Regulatory Agency (MHRA); UK Government. *Impact of AI on the Regulation of Medical Products*; UK Government: London, UK, 2024. Available online: https://assets.publishing.service.gov.uk/media/662fce1e9e82181baa98a988/MHRA_Impact-of-AI-on-the-regulation-of-medical-products.pdf (accessed on 24 March 2026).
119. Brione, P.; Gajjar, D. Artificial Intelligence: Ethics, Governance and Regulation. 2024. Available online: <https://post.parliament.uk/artificial-intelligence-ethics-governance-and-regulation/> (accessed on 24 March 2026).

120. Naik, N.; Hameed, B.M.Z.; Shetty, D.K.; Swain, D.; Shah, M.; Paul, R.; Aggarwal, K.; Ibrahim, S.; Patil, V.; Smriti, K.; et al. Legal and Ethical Consideration in Artificial Intelligence in Healthcare: Who Takes Responsibility? *Front. Surg.* **2022**, *9*, 862322. [[CrossRef](#)] [[PubMed](#)]
121. Natali, C.; Marconi, L.; Dias, D.; Cabitza, F. AI-induced Deskillling in Medicine: A Mixed-Method Review and Research Agenda for Healthcare and Beyond. *Artif. Intell. Rev.* **2025**, *58*, 356. [[CrossRef](#)]
122. Bollaert, M.; Augereau, O.; Coppin, G. Measuring and Calibrating Trust in Artificial Intelligence. In *IFIP Conference on Human-Computer Interaction*; Springer Nature: Cham, Switzerland, 2024.
123. Morales-Forero, A.; Rueda Jaime, L.; Gil-Quñones, S.R.; Barrera Montañez, M.Y.; Bassetto, S.; Coatanea, E. An insight into racial bias in dermoscopy repositories: A HAM10000 data set analysis. *JEADV Clin. Pract.* **2024**, *3*, 836–843. [[CrossRef](#)]
124. Liopyris, K.; Gregoriou, S.; Dias, J.; Stratigos, A.J. Artificial Intelligence in Dermatology: Challenges and Perspectives. *Dermatol. Ther.* **2022**, *12*, 2637–2651. [[CrossRef](#)] [[PubMed](#)]

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.