

Research on an Explainable Artificial Intelligence Framework of Interpretable and Robust Skin Lesion Classification with Deep Learning Review

Rahul Agnihotri

agnihotrirahul2006@gmail.com

Dr. Abhay Shukla

abhayshukl@gmail.com

CSE Department, Rama University, Kanpur

Abstract- Skin cancer, in particular, melanoma is an important international problem since it is ranked among the most morbid and fatal diseases particularly when the disease is diagnosed in its advanced stages. Over the past years, deep learning (DL) models have shown impressive results when applied to the automatic classification of dermoscopic-image skin lesions. Nonetheless, the majority of the state-of-the-art DL models are black-box, and they restrict their transparency, clinical trustworthiness, and practical use. This review paper includes an in-depth and systematic analysis of the development of explainable artificial intelligence (XAI) methods towards obtaining interpretable and robust skin lesion classification with deep learning technologies. It also compares the traditional ways of machine learning and the nowadays methods using convolutional neural networks and outlines the advantages and disadvantages of the two types of techniques concerning interpretability, robustness, and clinical use. The specific focus is made on methods of intrinsic and post-hoc explainability, such as saliency-based visualizations, class activation mapping, gradient-based attribution methods, concept-based explanations, and architectures of interpretable models. The review also examines the enhancement strategies of robustness, including data augmentation, adversarial learning, uncertainty estimation, and model calibration in order to enhance the reliability of skin lesion classifier operating under real-world environments. Popular evaluation measures, benchmark data sets that are publicly accessible, and clinical validation guidelines are also critically scrutinized. In this paper, a combined XAI-based system is suggested that summarizes the latest developments, formulates the current problems, and exposes the research gaps related to the area. The framework should generate high diagnostic accuracy and offer clinically significant explanations through synthesizing existing knowledge. This review is supposed to assist researchers and practitioners in creating transparent, reliable and clinically deployable deep learning systems to automatically analyze skin lesions, which in the end can help the dermatologists make clinical decisions and achieve better patient outcomes.

Keywords- Explainable artificial intelligence, skin lesion classification, deep learning, interpretability, robustness, dermoscopic image analysis, medical image diagnosis

Introduction

Skin-related diseases are a major health problem in all parts of the world with skin cancer being among the most common and fastest growing malignancy. Of all its manifestations, melanoma is especially aggressive and causes a disproportionately large number of deaths caused by skin cancer even though it constitutes a smaller part of the overall cases. Proper and timely diagnosis is thus very vital in enhancing the survival of patients and lowering the levels of complexity in treatment [7]. Traditionally, the diagnosis of skin lesions is based on visual examination, dermoscopic analysis and histopathology of the case analyzed by skilled dermatologists. These methods are however time consuming, subjective, overly reliant on clinical judgment, and may differ among practitioners and in diverse healthcare environments [3]. The need to overcome such limitations has led to the creation of computer-aided diagnosis (CAD) systems that can help dermatologists in clinical decision-making.

As medical imaging and computational power technologies have improved, artificial intelligence (AI) has become a disruptive technology to use in automated skin lesion detection and classification with a specific focus on machine learning (ML) and deep learning (DL) [12]. CNNs and their variations have shown good results in recognizing benign and malignant lesions based on dermoscopic images, with a high accuracy rate similar to the rate of highly trained clinicians in the same case [1]. In contrast to traditional ones, DL-based models learn hierarchical feature representations directly on raw image data and do not require the use of manual feature engineering nor support end-to-end learning networks. Consequently, the high classification accuracy, sensitivity, and specificity of many studies have been found on standard benchmark datasets [9].

Even with these promising results, few studies have been done to translate deep learning models into practical clinical use. One of the most significant obstacles is the fact that most DL architectures are not transparent and they are done

10.48047/jocaaa.2024.33.08.478

as black-box models and do not give much information about how their decisions are made [6]. Healthcare professionals who make diagnostic choices in clinical areas where patient safety and medical treatment are directly affected by these choices need not only precise predictions, but also clear and interpretable explanations that are consistent with known medical knowledge. This lack of explainability may undermine the trust of the clinicians, make it difficult to pass the regulatory approval and raise the ethical and legal issues associated with accountability and algorithmic bias [14].

In order to overcome these issues, explainable artificial intelligence (XAI) has become a steadily growing focus as a solution to close the knowledge gap between high-performing and deep learning models and the interpretability requirements in clinical settings. XAI seeks to improve the transparency of models by allowing human interpretations of the way AI systems learn, act and make predictions [4]. Regarding the skin lesion classification, XAI methods can emphasize the areas of the image that influence the model decision, measure the significance of the learned features, and infer predictions based on clinically relevant attributes of asymmetry, border irregularity, color variation, and texture patterns [11]. Such exegeses can help dermatologists to authenticate AI outputs, diagnose possible mistakes, and appropriately incorporate AI-based information in their clinical workflow.

The currently used XAI methods that are used to analyze skin lesions can be broadly divided into the post-hoc and intrinsic explainability methods. Saliency map and other post-hoc methods, such as Gradient-weighted Class Activation Mapping (Grad-CAM), Layer-wise Relevance Propagation (LRP), and SHAP-based visualizations, do not change the model structure [2]. Although these approaches are quite popular because of their flexibility and ease of integration, they can be characterized by instability, low faithfulness, and noise sensitivity. By contrast, intrinsic or interpretable models are models that aim to be explainable, and include features that are understandable like attention module, prototype based learning or concept bottleneck architectures [8]. Whereas such methods are more likely to give more uniform explanations, they usually imply trade-offs between predictive performance and interpretability.

In addition to interpretability, the AI systems that should be applied in the clinical setting should be robust. In practice, the dermoscopy appearances can differ significantly because of the variations in light, the photography equipment, the skin pigmentation, the lesion sizes, and the availability of artifacts (e.g., hair, air bubbles) [10]. Deep learning models that are trained on few or biased data can not be very useful when they come across such variations and hence unreliable predictions. To address those problems, a number of robustness-improving measures, including data augmentation, domain adaptation, adversarial training,

uncertainty estimation, and ensemble learning, have been investigated [5]. Nevertheless, robustness and explainability are typically examined separately, and their joint influence on the creation of trustful clinical AI systems is not well examined.

The use of AI-based medical decision-support systems is also influenced by ethical and regulatory issues. The regulatory authorities are putting more focus on transparency, fairness, and accountability in healthcare AI-based solutions [15]. Explainable AI can be essential in satisfying these needs by making it possible to audit it, identifying bias, and having informed consent. XAI may be used in skin lesion classification to detect spurious dependencies, biases in the dataset, and overfitting to non-clinical artifacts, which improves the ethical impartiality and clinical integrity of AI models [13].

Considering the booming research in this field, it is increasingly necessary to have a systematic review, which would thoroughly investigate explainable and robust deep learning methods to classify skin lesions. Available surveys concentrate either on performance based deep learning architectures or general XAI algorithms with little discussion on dermatological applicability and application in the clinical setting [6]. An integrated view incorporating interpretability, robustness, considering data sets and evaluation practices is required to inform the future research and make a practical use of it.

To fill these gaps, this review article will provide a detailed discussion of the current developments in explainable and robust deep learning frameworks to classify skin lesions. It looks at popular datasets, architecture of models, and methodologies of explainability, strategies of robustness, evaluation metrics, and outlines current drawbacks and potential areas of future research. This work aims to contribute to the creation of open, responsible, and clinically viable AI systems that help dermatologists and eventually positively transform patient outcomes by compiling the existing knowledge and identifying the main issues [2].

Literature Review

The emergence of recent developments in the field of deep learning has radically altered the automated classification of skin lesions, providing a high detection limit with the help of both the dermoscopic and clinical images. Nonetheless, the lack of transparency of deep neural networks has prompted serious questions of trust, interpretability and clinical adoption. In turn, explainable artificial intelligence (XAI) has become one of the most important research directions, which has the role of creating interpretable, reliable, and robust decision-support systems in dermatological diagnosis [1], [5], [15]. This area provides a literature review on the available data on the explainable deep learning methods in skin lesion and skin cancer classification, focusing on the methodological tendencies,

10.48047/jocaaa.2024.33.08.478

the methods of explainability, the problem of robustness, and the limitations that are identified.

Initial studies were mainly aimed at achieving the best classification accuracy with the use of convolutional neural networks (CNNs), and little effort was made towards the interpretability need [18]. Following the stagnation in model performance, scientists started using more and more post-hoc explanation methods to present a visualization of model attention and elucidate decision-making mechanisms. The saliency maps, Grad-CAM, LIME and SHAP-based approaches have been extensively used to find areas in the lesions that impact prediction, hence enhancing clinician confidence and model validation [3], [11], [21]. Although these studies reported consistency between explanation maps and clinically relevant lesion features including asymmetry, boundary irregularity and color variation, issues regarding inconsistency and sensitivity to noise still remain [8].

A number of systematic reviews have pointed out that the explainability in skin cancer detection is not restricted to visualization but rather includes wider aspects of requirements like clinical relevance, explanation faithfulness, and usability [1]. Evidence-based research has also shown that XAI methods can be utilized not only to classify lesions but also to predict the risk of other conditions by using facial or skin images, which suggests that explainable systems can be expanded to other areas of dermatology [2]. These attempts enhance the significance of interpretability in preventive and personalized medicine, as well as touching upon the ethical issues of bias and demographic sensitivity [13].

There have been studies where hybrid methods which combine deep learning with conventional machine learning classifiers and optimization methods to improve interpretability and robustness have been investigated [4], [27]. Transparency gained on a feature level based on handcrafted descriptors or CNN-optimized embeddings has been demonstrated to be useful; but these methods can usually compromise end-to-end learning pipelines [3].

More lately, transformer-based architectures have become an area of interest because they can learn global contextual relationship in images [6], [30]. Vision transformers have shown competitive results when used with XAI techniques, better performance on localizing attention, and of interest in multi-class classification of skin lesions. Although these benefits are evident, issues of computational complexity and the consistency of explanation across layers are yet to be solved [10]. The comparison studies often suggest that hybrid CNN-Transformer networks are more effective in classification and interpretation compared to separate models [17].

Another requirement of a reliable skin lesion classification system has also been found to be robustness. Image capture

variability, class variance, and bias in the datasets are important factors that influence the model generalization abilities [24]. To solve the problems, some studies have integrated robustness-enhancing techniques like data augmentation, ensemble models, and uncertainty assessment, and federated learning into XAI-enabled systems [19], [27]. The explainable models, in particular the federated ones, have proven to be promising as they also maintain privacy of data as well as allow interpretation of data, which is applicable in distributed health care setting [19].

The issues of class imbalance, the lack of representation of rare types of lesions are the ongoing problems of diagnosis of skin cancer. Recent reviews point out that explainable models are not only expected to work well on majority classes but also provide meaningful explanations of minority classes to prevent misleading clinical interpretations [24]. Other hybrid CNNXAI systems have tackled this problem by using mechanisms of fairness-aware learning and class-sensitive explanation methods [13].

On top of image-level explanations, recent studies have investigated the biomarker discovery and concept-level interpretability with XAI methods [20]. The goals of these methods are to overcome the disconnect between data-driven predictions and the biomedical knowledge that allows carrying out translational research and informed clinical reasoning. In the same vein, rule-based fuzzy logic augmented with deep learning systems have also been suggested to increase human interpretability and decision transparency [29].

Nevertheless, there are a number of restrictions despite the high level of advancement. Post-hoc explanatory procedures are not always faithful, explainability is not often tested with standardized quantitative measures and few studies have been rigorously clinically validated [8], [15]. Moreover, there are no generally accepted criteria of explainability evaluation which makes the comparative analysis of various methods difficult [12]. The latest conference papers and preprint research are increasingly promoting unified frameworks that can work together to maximize the accuracy, robustness, fairness, and interpretability [16], [25], [26].

On the whole, the literature suggests a definite solution of accuracy-based models in favor of explainable, robust, and ethically justified AI models in the area of skin lesion classification. However, more development is needed to develop standardized evaluation criteria, combine domain-specific medical experience, and foster clinical implementation. There is a need to establish a cohesive XAI framework where predictive performance is balanced to provide meaningful and clinically relevant explanations to translate research progress in routine dermatological practice [22], [30].

Table 1: Overview of Explainability Techniques Used in Skin Lesion Classification Studies

Explainability Approach	Description	Commonly Used With	Key Strengths	Limitations
Saliency / Grad-CAM	Gradient-derived visual maps that emphasize image regions influencing model predictions	CNN, ResNet, DenseNet	Simple to apply and visually intuitive	Prone to noise and instability
LIME / SHAP	Local interpretable surrogate models that estimate feature contributions	CNN, hybrid ML-DL models	Model-agnostic and highly flexible	Limited spatial accuracy for images
Attention mechanisms	Internally learned spatial focus integrated within the network architecture	CNN, Transformers	Built-in interpretability during training	Increased architectural and computational complexity
Prototype / Concept-based	Classification decisions guided by learned representative visual concepts	CNN, hybrid models	Produces explanations aligned with clinical reasoning	Lower adaptability and reduced model flexibility

Table 2: Architectural Trends in Explainable Skin Lesion Classification

Architecture Type	Explainability Integration	Reported Advantages	Key Challenges
CNN-based models	Post-hoc XAI techniques (e.g., Grad-CAM, LIME)	High classification accuracy and structural simplicity	Core decision process remains a black box
Transformer-based	Visualization of attention	Effective modeling of	High computational

Architecture Type	Explainability Integration	Reported Advantages	Key Challenges
models	weights	global image context	and memory requirements
Hybrid CNN-ML models	Feature-level interpretability methods	Enhanced transparency of decision-making	Dependence on multi-stage processing pipelines
Federated models	Explainability in distributed learning frameworks	Privacy-aware and decentralized model training	Increased communication and synchronization overhead

Table 3: Major Research Gaps Identified in Existing Literature

Aspect	Identified Gap	Research Implication
Explainability evaluation	Absence of standardized quantitative assessment metrics	Limits objective comparison between models
Clinical validation	Minimal involvement of practicing dermatologists	Low confidence in real-world clinical deployment
Robustness	High sensitivity to noise, bias, and data variability	Poor generalization across diverse clinical settings
Standardization	Lack of a unified XAI framework	Leads to fragmented and inconsistent research outcomes

The literature review shows that explainable deep learning models on skin lesion classification have improved swiftly, but further studies need to focus on creating combined, clinically valid, and effective XAI models to make them safe and reliable during their application in a healthcare facility.

Proposed Methodology

The methodological development of explainable artificial intelligences (XAI) to classify skin lesions is a slow shift of the performance based deep learning models to models that are transparent, robust and clinically relevant. The scope of methods used in existing research is quite broad and can be roughly grouped into four main areas of analysis, namely the model architectures, techniques of explainability integration, methods of robustness enhancement, and evaluation protocols. This part analyzes these methodological decisions critically on the basis of recent studies that have been reported in the literature [1] -[30].

Deep Learning Architectures for Skin Lesion Classification

The majority of the available approaches are based on convolutional neural networks (CNNs), due to their ability to extract spatial and texture-related characteristics on the dermoscopic images [5], [9], [18]. The popular architectures include ResNet, DenseNet, EfficientNet, and CNN variants with specific customizations, which are usually trained with the help of transfer learning to address the problem of a limited supply of annotated medical imaging data [11], [14]. These models are effective, though they are not necessarily transparent and therefore have stimulated the incorporation of elements of explainability mechanisms either by adding new architectural layers or by employing post-hoc methods of explanation.

Table 4: Comparative Analysis of Model Architectures Used in Skin Lesion Classification

Architecture Type	Learning Strategy	Interpretability Level	Key Strengths	Limitations	References
CNN-based	End-to-end supervised learning	Low (intrinsic interpretability)	High predictive accuracy and availability of mature development tools	Decision-making remains largely opaque	[5], [9], [18]
Transformer-based	Attention-driven learning	Medium	Effective capture of global contextual information	Requires large datasets and high computational resources	[6], [30]
Hybrid DL-ML	Two-stage or modular learning	Medium to High	Enhanced transparency and explainability	Increased system complexity and pipeline overhead	[3], [23], [27]
Federated DL	Distributed collaborative learning	Medium	Enables privacy-preserving model training	Communication and synchronization overhead	[19]

10.48047/jocaaa.2024.33.08.478

More recent works have begun considering the use of Vision Transformers (ViTs) and CNN + Transformer models to learn long-range dependencies and global contextual information in skin lesion images [6], [30]. Methodologically, these models are more interpretable (particularly with respect to self-attention), since the weights of attention can be visualized to reveal influential regions of images. Although this has these advantages, the approaches based on transformers typically inquire significantly larger training datasets and more computational resources, limiting their use in large-scale clinical settings [10], [17].

The combination of deep learning with the feature extractor and traditional machine learning classification methods, such as support vectors machine (SVMs), Random Forests, or Radial Basis Function Networks, has also been investigated [3], [23], [27]. These ideas are used to achieve equilibrium between high representational learning during the feature extraction phase, and better interpretability during the decision-making layer. Although the additional transparency contributes to the explainability of the model, hybrid systems in general can jeopardize end-to-end optimization and add more complexity to the training and deployment pipelines.

Explainability Integration Strategies

The methods of explainability used in the classification of skin lesions can be divided into post-hoc and intrinsic broadly. The prevailing literature relies on the use of post-hoc explainability methods because these are model agnostic. Grad-CAM, LIME, SHAP, and saliency maps are the most common techniques used to visualize or feature-based explain model predictions applied to trained CNNs [1], [5], [21]. These approaches do not necessitate any adjustment of the initial model architecture, are easy to incorporate, and these approaches are effectively applicable to retrospective interpretability analysis. Nevertheless, post-hoc explanations are not necessarily an accurate reflection of the reasoning that transpired behind the scenes of the model, which can result in its misunderstanding within the clinical setting [8], [15]. There are also a number of studies that have documented instability in explanation maps with regard to minor perturbation of input casting doubts on the reliability and consistencies of such visualizations [12].

Table 5: Overview of Explainability Methodologies and Their Key Characteristics

Explainability Type	Representative Techniques	Stage of Integration	Key Advantages	Methodological Limitations	References
Post-hoc	Grad-CAM, LIME, SHAP	Applied after model training	Model-agnostic and easy to integrate	Potential lack of faithfulness to true model	[1], [5], [21]

10.48047/jocaaa.2024.33.08.478

Explainability Type	Representative Techniques	Stage of Integration	Key Advantages	Methodological Limitations	References
				reasoning	
Intrinsic	Attention mechanisms, prototype-based models	Embedded during model training	Produces stable and consistent explanations	Imposes architectural design constraints	[14], [28]
Feature-based	Handcrafted or pre-extracted descriptors	Input or feature representation level	Explanations align with human-understandable features	Disrupts end-to-end learning frameworks	[3], [11]
Optimization-based	Particle Swarm Optimization, evolutionary XAI methods	Integrated during training optimization	Enables joint optimization of accuracy and explainability	High computational complexity	[4]

As a solution to these problems, intrinsic explainability methods incorporate attention modules, prototype learning or concept bottleneck layers directly into the network architecture [14], [28]. Such methods tend to offer better and more accurate explanations; though can usually limit architectural adaptability, resulting in less adaptability and performance inertia. One more avenue of explainability has been via optimization based approaches, where the weights of features used in feature-selection or the attention parameters are optimized by evolutionary algorithms or swarm-based techniques [4]. Although these approaches can be used to combine interpretability and predictive performance, they substantially raise the cost of computations and the complexity of tuning.

Robustness and Generalization Methodologies

There are two broad methods of methodological approaches to enhance robustness, namely data-centric and model-centric approaches. The use of data augmentation is still one of the most popular methods used to enhance the generalization of models in different lighting conditions, noise intensities, and imaging set-ups [9], [18]. Other studies have also used superior augmentation techniques and rebalancing of classes to overcome severe class imbalance that is usually present in skin lesion data [24].

Model-centric methods of robustness such as ensemble learning, uncertainty estimation, and adversarial training, are all designed to minimize model overconfidence and improve the predictive accuracy of models [15], [27]. Federated learning models also make significant contributions to the robustness of that they allow decentralized model training to be performed by more than one institution without violating patient privacy. Nonetheless, these models bring about new issues to do with model aggregation and consistency of the generated explanation using distributed environments [19].

Even with such developments, robustness and explainability have traditionally been considered as different goals in the current literature. There have been few studies that examined the stability of explanations with noisy or adversarial conditions, which is an important research gap in the construction of reliable clinical AI systems [8], [12].

Table 6: Robustness-Oriented Methodological Strategies in Skin Lesion Classification

Robustness Strategy	Level of Implementation	Primary Target	Impact on Explainability	References
Data augmentation	Data preprocessing stage	Overfitting and dataset bias	Neutral to slightly positive influence	[9], [18]
Ensemble learning	Model-level integration	Reduction of prediction variance	Decreased clarity of individual explanations	[27]
Uncertainty estimation	Output or decision level	Mitigation of model overconfidence	Enhances interpretability and user trust	[15]
Federated learning	System-level deployment	Data privacy and scalability	Challenges in aggregating consistent explanations	[19]
Fairness-aware learning	Training objective formulation	Bias reduction across demographic groups	Supports ethical and socially responsible interpretability	[13], [24]

Evaluation Protocols and Methodological Limitations

The standard classification metrics used in the evaluation practices of literature are mostly accuracy, sensitivity, specificity, F1-score and the area under the curve (AUC)

[5], [9], [18]. Nonetheless, the explainability is evaluated mostly by qualitative means (often by visual inspection of heatmaps or attention maps) and scarce quantitative validation techniques are used [8]. Additionally, very few of the studies include dermatologists or clinical experts in the evaluation process, which undermines the clinical relevance and reliability of the produced explanations [1], [15].

The lack of standard messages to compare the explainable AI (XAI) techniques is another important limitation of the methodology. Comparisons across studies are not easy to achieve because of the number of differences in datasets, methods of explainability, and scales of visualization used in varied works [12], [22]. In order to deal with this problem, more recent studies have been advancing in the position of multidimensional evaluation models that jointly evaluate predictive performance, as well as, explanation faithfulness, robustness, and fairness [25], [30].

In a wider sense, the discipline is moving away, though, toward explainable, robust, and ethically-friendly skin lesion classification AI systems, which are based on conventional black-box deep learning models. Although convolutional neural networks and transformer-based architectures are the most popular at the moment, methods of explaining the nature of architectures and enhancing their robustness are disjointed. The main issue in the methodology is the absence of a single set of frameworks and standardized evaluation procedures that would all maximize accuracy, interpretability and robustness. To open up these blind spots, it is crucial to have the capacity to translate XAI-enabled dermatological decision-support systems to rely on clinical translation.

Conclusion and Future Scope

This paper has reviewed critically the developments of explainable artificial intelligence (XAI) methods to attain interpretable and robust skin lesion classification with deep learning technologies. The summary of the analysis shows that, although convolutional neural networks and the upcoming transformer-based models have proven themselves to be highly diagnostic, the black-box property remains the barrier to the widespread use of these models in clinical settings since issues of trust, transparency, and scalability are still of concern. Therefore, explainability has been shown to become a mandatory part already instead of a non-compulsory improvement in state-of-the-art skin cancer diagnostics.

According to the reviewed methodologies, the most popular methodologies are now post-hoc explainability techniques, including saliency maps, Grad-CAM, LIME, and SHAP, because of their model-agnostic nature and simple to apply. Nevertheless, they are prone to instability and lack of faithfulness which highlights the necessity of more inherent and conceptual explainability mechanisms. Attention modules, prototype-based learning and interpretable

10.48047/jocaaa.2024.33.08.478

bottlenecks have been demonstrated to be useful in producing a more reliable and clinically meaningful explanation, although at the cost of increased architectural complexity or computational complexity.

Strength and generalization are other important issues, especially during discrepancies in imaging parameters, disproportion of classes and demographic biasness in datasets. Though a number of measures have been pursued to solve these problems (data augmentation, ensemble learning, the uncertainty estimation and federated learning), they have not been integrated with the mechanism of explainability. Recent evidence proposes that optimization of robustness and interpretability should be combined to produce credible and trustful clinical decision support systems.

All in all, it is evident in the literature that there are no standardized evaluation protocols on how to measure the quality, reliability, and clinical relevance of explainability. Future studies ought to be done in the creation of single XAI frameworks that favor diagnostic performance with reliable, consistent, and equitable elaborations, with clinician-in-the-loop assessment. These developments are essential to the translation of explainable deep learning models based on experimental research studies into safe, ethical, and useful tools of everyday skin cancer detection and better patient care.

References

- [1] A Hauser, Katja, Alexander Kurz, Sarah Haggenueller, Roman C. Maron, et al. "Explainable Artificial Intelligence in Skin Cancer Recognition: A Systematic Review." *International Journal of Cancer* 151, no. 6 (2022): 793–812. <https://doi.org/10.1002/ijc.33909>. ISSN: 0020-7136.
- [2] Liu, Xin, Thijs E. Sangers, Tamar Nijsten, Manfred Kayser, Luis M. Pardo, et al. "Predicting Skin Cancer Risk from Facial Images with an Explainable Artificial Intelligence (XAI) Based Approach: A Proof-of-Concept Study." *The Lancet Digital Health* 6, no. 3 (2024): e182–e190. [https://doi.org/10.1016/S2589-7500\(23\)00244-3](https://doi.org/10.1016/S2589-7500(23)00244-3). ISSN: 2589-7500.
- [3] Khater, Tarek, Shadab Ansari, Sherif Mahmoud, Abir Hussain, et al. "Skin Cancer Classification Using Explainable Artificial Intelligence on Pre-Extracted Image Features." *Intelligent Systems with Applications* 18 (2023): 100224. <https://doi.org/10.1016/j.iswa.2023.100224>. ISSN: 2667-3053.
- [4] Shah, Syed Abdul Hannan, Syed Tahir Hussain Shah, Rabia Khaled, et al. "Explainable AI-Based Skin Cancer Detection Using CNN, Particle Swarm Optimization and Machine Learning." *Journal of Imaging* 10, no. 1 (2024): 1–18. <https://doi.org/10.3390/jimaging10010001>. ISSN: 2313-433X.
- [5] Nigar, Nazia, Muhammad Umar, Muhammad Khalid Shahzad, Saif Islam, and David Abalo. "A Deep Learning Approach Based on Explainable Artificial Intelligence for Skin Lesion Classification." *IEEE Access* 10 (2022): 71544–71560. <https://doi.org/10.1109/ACCESS.2022.3188765>. ISSN: 2169-3536.
- [6] Dagnaw, Getaneh Hailu, Mohammed El Mouhtadi, and Sanaa El Filali. "Skin Cancer Classification Using Vision Transformers and Explainable Artificial Intelligence." *Journal of Medical Artificial Intelligence* 7 (2024): 1–12. <https://doi.org/10.21037/jmai-23-45>. ISSN: [Not provided].

10.48047/jocaaa.2024.33.08.478

- [7] Ramoliya, Fenil, Khyati Gohil, Aayush Gohil, Rishabh Gupta, et al. "X-CaD: Explainable AI for Skin Cancer Diagnosis in Healthcare 4.0 Telesurgery." *IEEE International Conference on Communications (ICC)*, 2024. ISSN: [Conference paper].
- [8] Sangwan, Harsh. "Quantifying Explainable AI Methods in Medical Diagnosis: A Study in Skin Cancer." *medRxiv*, 2024. <https://doi.org/10.1101/2024.03.15.24304321>. ISSN: [Preprint].
- [9] Ahmad, Naveed, Jawad Hussain Shah, Muhammad Attique Khan, Javed Baili, et al. "A Novel Framework of Multiclass Skin Lesion Recognition from Dermoscopic Images Using Deep Learning and Explainable AI." *Frontiers in Oncology* 13 (2023): 1–18. <https://doi.org/10.3389/fonc.2023.1170024>. ISSN: 2234-943X.
- [10] Radiah, Farhana, Kazi Rahman, Lamia Asadullah, Mohammad Saifur Rahman Sohan, et al. "Explainable AI (XAI) Driven Skin Cancer Detection Using Transformer and CNN Based Architecture." *BRAC University Institutional Repository*, 2023. ISSN: [Thesis].
- [11] Mridha, Kabir, Md. Milon Uddin, Jooyoung Shin, Samir Khadka, et al. "An Interpretable Skin Cancer Classification Using Optimized Convolutional Neural Network for a Smart Healthcare System." *IEEE Access* 11 (2023): 41003–41018. <https://doi.org/10.1109/ACCESS.2023.3269991>. ISSN: 2169-3536.
- [12] Azlaan, Hira, and Joshua Oluwaseyi. "Explainable AI (XAI) for Skin Cancer Detection." *ResearchGate*, 2024. ISSN: [Preprint].
- [13] Veeramani, Naveen, and Premalatha Jayaraman. "YOLOv7-XAI: Multi-Class Skin Lesion Diagnosis Using Explainable AI With Fair Decision Making." *International Journal of Imaging Systems and Technology* 34, no. 1 (2024): e23012. <https://doi.org/10.1002/ima.23012>. ISSN: 0899-9457.
- [14] Hamim, Syed Abrar, Md. Umar Iqbal Tamim, Muhammad Fahim Mridha, et al. "SmartSkin-XAI: An Interpretable Deep Learning Approach for Enhanced Skin Cancer Diagnosis in Smart Healthcare." *Diagnostics* 14, no. 2 (2024): 1–20. <https://doi.org/10.3390/diagnostics14020189>. ISSN: 2075-4418.
- [15] Alrabai, Abdullah, Aymen Echtioui, and Fares Kallel. "Explainable Deep Learning Approaches for Skin Cancer Diagnosis." *Network Modeling Analysis in Health Informatics and Bioinformatics* 14, no. 1 (2025): 1–15. ISSN: 2192-6662.
- [16] Thapar, Priyanka, and Shailendra Tiwari. "Empowering Skin Cancer Diagnosis: Integrating Advanced Deep Learning Models with Explainable AI for Lesion Classification." *IEEE International Conference on Computing*, 2024. ISSN: [Conference paper].
- [17] Sree, N. Harish, K. Trisha, P. Sai Reddy, et al. "Explainable AI Insights into Skin Cancer Detection: A Comparative Study of CNN, DenseNet, and ResNet." *IEEE International Conference on Artificial Intelligence*, 2024. ISSN: [Conference paper].
- [18] Rehman, Muhammad Zia Ur, Fawad Ahmed, Shaha Alsuhbany, et al. "Classification of Skin Cancer Lesions Using Explainable Deep Learning." *Sensors* 22, no. 5 (2022): 1–18. <https://doi.org/10.3390/s22051823>. ISSN: 1424-8220.
- [19] Biswas, Saptarsi, Soham Saha, Md. Sharif Uddin, and Rafia Mostafiz. "An Explainable and Federated Deep Learning Framework for Skin Cancer Diagnosis." *PLoS ONE* 20, no. 1 (2025): e0294321. <https://doi.org/10.1371/journal.pone.0294321>. ISSN: 1932-6203.
- [20] Meena, Jai, and Yogesh Hasija. "Application of Explainable Artificial Intelligence in the Identification of Squamous Cell Carcinoma Biomarkers." *Computers in Biology and Medicine* 145 (2022): 105393. <https://doi.org/10.1016/j.compbiomed.2022.105393>. ISSN: 0010-4825.
- [21] Gupta, Rajesh Kumar. "Interpretable AI-Enabled Model for Skin Cancer Diagnosis Using LIME." *Procedia Computer Science* 225 (2025): 1234–1243. <https://doi.org/10.1016/j.procs.2025.05.012>. ISSN: 1877-0509.
- [22] Alrabai, Abdullah, Aymen Echtioui, and Fares Kallel. "Explainable Pre-Trained Models for Skin Cancer Classification." *IEEE International Conference on Machine Learning*, 2025. ISSN: [Conference paper].
- [23] Ullah, Md. Asad, and Tariq Zia. "Hybrid Interpretable Deep Learning Framework for Skin Cancer Diagnosis: Integrating Radial Basis Function Networks with Explainable AI." *arXiv preprint arXiv:2501.14885*, 2025. ISSN: [Preprint].
- [24] Farooq, Ayesha Ben, Sheraz Akbar, Qurat ul Ain, et al. "Skin Cancer Detection and Classification Using Explainable Artificial Intelligence for Unbalanced Data: State of the Art." *Explainable Artificial Intelligence in Medical Decision-Making Systems*, 2025. ISSN: [Book chapter].
- [25] Kumar, Manish Kumar, J. Vaishnavi, T. Anjibabu, et al. "Skin Cancer Detection: A Hybrid Approach Combining CNNs and Explainable AI." *IEEE Conference on Machine Learning Applications*, 2025. ISSN: [Conference paper].
- [26] Aftab, Rida Saeed, Syed Abrar Hamim, Md. Mahmud Ahmed, et al. "SkinScanNet: A CNN-Based Model with Explainable AI for Reliable and Transparent Skin Cancer Detection." *IEEE Conference on Computer and Communication*, 2024. ISSN: [Conference paper].
- [27] Roy, Anik Lal, Al Maruf, Rifat Islam Sony, et al. "Fostering Trust and Interpretability in Skin Cancer Classification: A Hybrid Framework of Deep Learning and Machine Learning with Explainable AI." *Journal of Ambient Intelligence and Humanized Computing* 16, no. 2 (2025): 1–15. ISSN: 1868-5137.
- [28] Munjal, Gaurav, Piyush Bhardwaj, Vikas Bhargava, et al. "SkinSage XAI: An Explainable Deep Learning Solution for Skin Lesion Diagnosis." *Health Care Science* 3, no. 1 (2024): 1–12. <https://doi.org/10.1002/hcs2.45>. ISSN: 2770-9019.
- [29] Lima, Sérgio, Luis Terán, and Edy Portmann. "A Proposal for an Explainable Fuzzy-Based Deep Learning System for Skin Cancer Prediction." *IEEE International Conference on Fuzzy Systems*, 2020. ISSN: [Conference paper].
- [30] Ali, Asad, Hamza Shahbaz, and Robertas Damaševičius. "xCViT: Improved Vision Transformer Network with Fusion of CNN and Xception for Skin Disease Recognition with Explainable AI." *Computers, Materials & Continua* 78, no. 1 (2025): 1–20. ISSN: 1546-2218