Studies in Medical and Health Sciences
https://smhs.sabapub.com
ISSN: 3007-3707

2025 Volume 2, Issue 1: 1– 14

DOI: https://doi.org/10.48185/smhs.v2i1.1422

Integrating Explainable AI for Skin Lesion Classifications: A Systematic Literature Review

Muhammad Bilal Jan[1], Muhammad Rashid[2], Raja Vavekanand[3*], Vijay Singh[4]

[1]University of Engineering and Technology, Peshawar, Pakistan
[2]Muhammad Nawaz Shareef University of Agriculture, Multan, Pakistan
[3]Datalink Research and Technology Lab, Islamkot, Pakistan
[4]Benazir Bhutto Shaheed University Lyari, Karachi, Pakistan

Received: 16.11.2024 • Accepted: 30.12.2024 • Published: 22.01.2025 • Final Version: 30.01.2025

Abstract: Skin cancer, particularly melanoma, poses a significant global health challenge due to its prevalence and mortality rate. Early detection is critical to improving outcomes, as advanced cases become increasingly difficult to treat. The advent of Artificial Intelligence (AI) and Explainable AI (XAI) techniques has revolutionized dermatological diagnostics by offering accurate and interpretable solutions. This systematic review investigates the integration of XAI in skin lesion classification, analyzing 22 recent studies published between 2019 and 2023. The studies encompass diverse approaches, including deep learning models like CNNs, ResNet, DenseNet, and MobileNet, as well as explainability techniques such as Grad-CAM, SHAP, and saliency maps. Results highlight significant advancements in accuracy and interpretability, with some models achieving over 99% accuracy on datasets like ISIC 2018 and HAM10000. However, challenges persist, including dataset imbalances, limited diversity in patient metadata, and generalizability across different skin types and imaging conditions. XAI methods, by visualizing model decision pathways, enhance transparency, fostering trust among clinicians and enabling seamless AI integration into clinical practice. This review underscores the potential of combining state-of-the-art AI models with explainable frameworks to address the complexities of skin lesion diagnostics. It advocates for future research to prioritize diverse, metadata-rich datasets, innovative optimization techniques, and robust architectures to develop reliable, interpretable diagnostic tools. By bridging the gap between advanced AI and user understanding, this work contributes to the creation of clinically applicable, trustable AI-driven healthcare solutions.

Keywords: Explainable AI, Skin Lesion, AI in Healthcare, Medical Imaging

1. Introduction

Skin lesions, ranging from benign growths to malignant cancers, represent a significant challenge in medical diagnostics, exacerbated by their diverse appearances and potential health implications (Ahmad et al., 2020; Ahmad et al., 2023; Ballari et al., 2022; Barata et al., 2020; Codella et al., 2019; Ding et al., 2023; El-Khatib et al., 2020). Melanoma and non-melanoma skin cancers are among the most prevalent and concerning types, necessitating accurate and timely diagnosis for effective treatment (Singh et al., 2024; Vavekanand, 2024). Skin cancers remain the most common group of diagnosed cancers across the globe, with an estimated 330,000 new cases of melanoma alone diagnosed globally (WHO, 2022). Early detection is key; new melanomas are shallower and thinner than those that have metastasized, decreasing treatment difficulty (Hoang et al., 2022; Howard et al., 2019; Iqbal

* Corresponding Author: bharwanivk@outlook.com

et al., 2020; Kassem et al., 2020; Kurasinski & Mihailescu, 2020; Lu & Li, 2020; Metta et al., 2021; Ni et al., 2021; Nigar et al., 2022; Olayah et al., 2023; Rehman et al., 2022).

Skin lesions can be classified into various categories (Figure 1). These include Melanoma (MEL), Melanocytic nevus (NV), Basal cell carcinoma (BCC), Actinic Keratosis (AK), Benign keratosis lesion (BKL), Dermatofibroma (DF), Vascular lesion (VASC), and Squamous cell carcinoma (SCC). Accurate classification is crucial for proper diagnosis and treatment, as each type of lesion has distinct characteristics and implications for patient care. Generally, skin cancers develop from sun exposure to ultraviolet (UV) rays, as well as artificial lighting from sunlamps and tanning beds. While the majority of skin cancers frequently occur and are easily treated, melanoma (MEL) constitutes over 70% of skin cancer fatalities despite representing around 5% of all skin cancers. With its propensity to grow and spread further than other types of skin cancers, its early detection is paramount to saving lives (El-Khatib et al., 2020; Nigar et al., 2022; Swamy & Divya, 2021; Tô et al., 2019; Tschandl et al., 2018a, 2018b; Vavekanand & Kumar, 2024; Vavekanand et al., 2024; Villa-Pulgarin et al., 2021). Challenges in accurately classifying skin lesions stem from their variability in appearance due to factors such as skin type, lesion morphology, and imaging conditions. Even trained dermatologists find difficulty in accurately diagnosing skin lesions, with only extensive training and experience leading to better diagnoses.

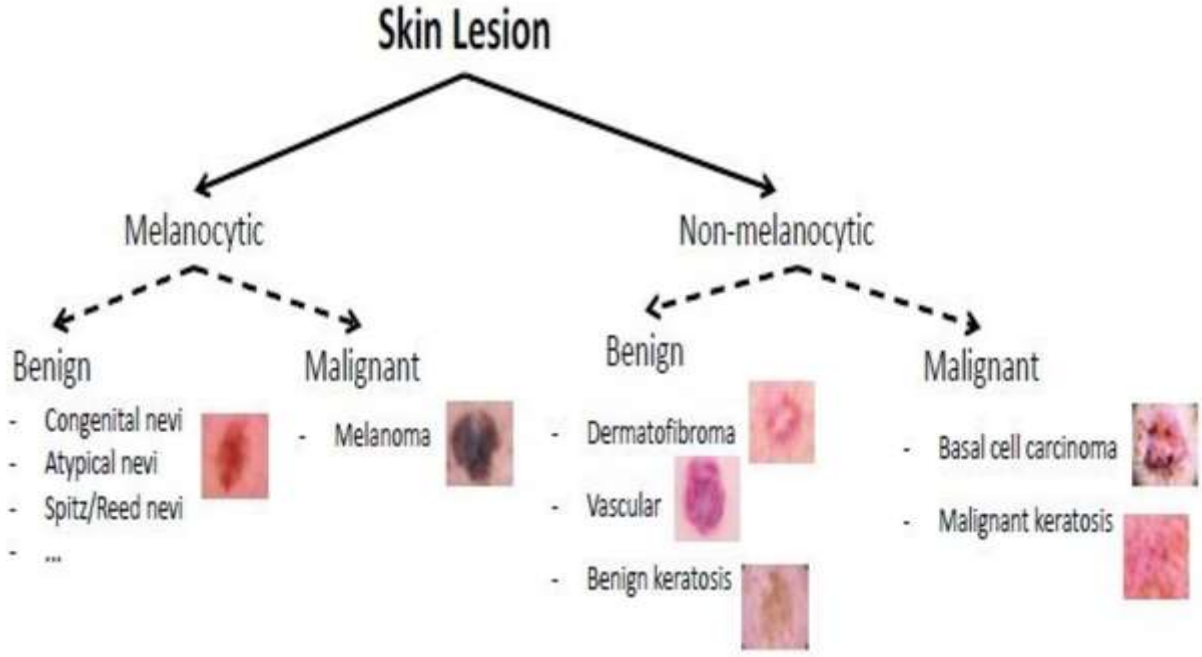

Fig 1. Skin lesions are categorized into melanocytic and non-melanocytic, each with benign or malignant subtypes, ultimately leading to a diagnosis of specific lesion types (Barata et al., 2019).

Achieving robust classification with artificial intelligence thus requires advanced feature extraction techniques and machine learning algorithms capable of discerning subtle differences crucial for diagnosis (Vavekanand & Kumar, 2024; Vavekanand et al., 2024). Explainable AI (XAI) methods play a pivotal role in enhancing transparency and interpretability in dermatological diagnostics due to the black-box nature of the diagnostics models. By employing techniques such as feature visualization, XAI elucidates the reasoning behind AI-driven diagnostic decisions. This transparency builds trust among healthcare providers and facilitates the integration of AI systems into clinical practice by empowering clinicians to understand and validate AI recommendations. Collaborative efforts, such as those supported by initiatives like ISIC, are crucial in advancing the field towards more reliable and accessible diagnostic tools for skin diseases (Lu & Li, 2020; Metta et al., 2021; Nigar et al., 2022; Rehman et al., 2022; Tschandl et al., 2018a, 2018b; Vavekanand & Kumar, 2024; Vavekanand et al., 2024; Villa-Pulgarin et al., 2021; Wu et al., 2019; Yang et al., 2019; Young et al., 2019; Barata et al., 2019).

This research report delves into the intersection of XAI and skin lesion classification to bridge the gap between highly advanced deep learning models and user understanding. By making clear the decision-making processes of the deep learning models with XAI, this report aspires to develop an accurate and explainable skin lesion classification system.

2. Literature Review

Since the inception of the International Skin Imaging Challenge (ISIC) in 2016, research on skin imaging and lesion classification has seen steady growth. The integration and accessibility of artificial intelligence (AI), combined with the ISIC competitions, have made this field increasingly approachable for researchers worldwide. With the growing emphasis on explainable AI (XAI), many studies have adopted explainability models in skin lesion classification. Explainable skin lesion classification, supported by a critical review discussion table to summarize findings, related works, and the background for the proposed method. We focus on three categories: general skin lesion classification studies, explainable skin lesion classification papers, and works specifically utilizing the ISIC 2019 dataset. All reviewed papers were published between 2019 and 2023.

2.1. Papers on Skin Lesion Classification

(Allugunti et al., 2022) classified three melanoma types—lesion malignant, superficial spreading, and nodular melanoma—using the DermNet dataset. Comparing decision trees, random forest, gradient-boosted trees, and CNN classifiers, the CNN model achieved the highest performance with precision (91.07%), recall (87.68%), F1-score (89.32%), and accuracy (88.83%). However, the study lacked details on the CNN architecture, feature representations, and experimental setup, limiting reproducibility. Future work should address these limitations.

(Ahmad et al., 2020), fine-tuned ResNet152 and InceptionResNet-V2 models with a triplet loss function to classify images from the AI Skin dataset. The models mapped input images to a 128-dimensional Euclidean space, comparing embeddings using L2 distances. The InceptionResNet-V2 + Triplet model outperformed others, achieving an accuracy of 87.42%, recall of 97.04%, and specificity of 96.48%. Despite surpassing prior benchmarks, the small dataset (800 images) with broad categories like acne and blackheads limited the study.

(Swamy & Divya, 2021) explored the effects of texture and color features on classification accuracy using decision tree and SVM models with the DermNet and DermQuest datasets. While decision tree models achieved accuracies of 66% (texture) and 75% (color), SVM models performed better with 75% (texture) and 83% (color). However, these results lagged behind state-of-the-art techniques, emphasizing the need for advanced feature extraction methods combining texture and color.

(Wu et al., 2019) used the Xiangya-Derm dataset with six common skin diseases, Wu et al. compared five CNN architectures: ResNet50, Inception V3, DenseNet-121, Xception, and Inception-ResNet-V2. Pre-trained models performed better than untrained ones, despite the dataset containing only facial skin images. InceptionResNet-V2 achieved the best precision (70.8%) and recall (77.0%). However, the dataset's limited size and lack of fine-tuning contributed to lower performance for specific classes like actinic keratosis (Table 1).

2.2. Explainable Skin Lesion Classification Papers

(Ahmad et al., 2023) utilized the ISIC 2018 and HAM10000 datasets, they selected features using the Butterfly Optimization Algorithm (IBOA) and fused them into Xception, ShuffleNet, and a fusion model. Grad-CAM visualizations highlighted prediction regions, and their model achieved accuracy (99.3%), recall (99.38%), precision (99.4%), and F1-score (99.38%). Despite high metrics, the study recommended optimization improvements, such as Bayesian optimization, for further refinement.

(Ballari et al., 2022) studied and employed ResNet18 with Grad-CAM on a Kaggle skin disease dataset, achieving 96% accuracy. The Grad-CAM outputs provided interpretable visualizations, but the study suffered from vague reporting, including an unspecified dataset size and unclear experimental details, limiting reproducibility.

(Barata et al., 2020) developed a hierarchical model inspired by dermatological decision-making for ISIC 2017 and 2018 datasets. Using an image encoder (VGG-16, ResNet-50, DenseNet-161), a hierarchical decoder, and an attention module, their model achieved sensitivity (86.7%), specificity (87.1%), and AUC (92.4%) with VGG-16. While the taxonomy improved accuracy, challenges remained in identifying melanoma and handling image transformations.

(Ding et al., 2023) model incorporated MobileViT blocks for improved classification accuracy and interpretability using Grad-CAM and AblationCAM. On the ISIC 2018 dataset, it achieved precision (93.1%), recall (93.2%), F1-score (93.1%), and accuracy (93.2%). Despite strong performance, the lack of diverse skin tones and patient indicators limited generalizability.

(Tschandl et al., 2018) used ResNet-50, Tschandl implemented Content-Based Image Retrieval (CBIR) on EDRA, ISIC 2017, and private datasets, achieving an accuracy of 76.2%, AUC (85%), specificity (92.2%), and sensitivity (72.7%). CBIR intuitively retrieved visually similar images but suffered from a lack of fine-tuning and dataset limitations.

(Young et al., 2019) combined KernelSHAP and Grad-CAM with Inception CNN to classify skin lesion images from a dataset of 6,017 images. The model achieved a mean AUC of 85% and a recall of 87%. Despite dual explainability techniques, dataset imbalances and reliance on Inception limited its potential.

(Rehman et al., 2022) compared MobileNetV2 and DenseNet201 with Grad-CAM visualization on an ISIC dataset. DenseNet201 achieved the highest metrics: accuracy (95.50%), precision (97.02%), F1-score (95.46%), sensitivity (95.96%), and specificity (97.06%). Further optimization methods could enhance the model's efficiency (Table 2).

2.3. ISIC 2019 Related papers

The International Skin Imaging Collaboration (ISIC) 2019 Challenge focused on classifying skin lesions into eight categories using image data, with or without patient metadata. This section summarizes key studies utilizing the ISIC 2019 dataset (Table 3).

2.3.1. Image-Only Classification

(El-Khatib et al., 2020) achieved 93% accuracy using GoogLeNet, ResNet-101, and NasNet-Large with a decision fusion model, but the small dataset size risked overfitting. Similarly, (Kassem et al., 2020) modified GoogLeNet, achieving 94.92% accuracy, outperforming the ISIC 2019 winning model. (Gong et al., 2020) integrated StyleGAN-generated images with ISIC 2019 data, achieving 99.5% accuracy using a fusion of 43 CNNs, addressing dataset imbalance but complicating model selection.

2.3.2. Incorporating Metadata

(Gessert et al., 2020) winners of ISIC 2019, used metadata with EfficientNet CNNs and a neural network, achieving an AUC of 98%. However, metadata inclusion reduced sensitivity. (Tô et al., 2019) combined UNet-based segmentation with EfficientNet-B4, though results were not reported.

2.3.3. Novel Approaches

(Hoang et al., 2022) proposed an entropy-based segmentation with Wide-ShuffleNet, yielding 82.56% accuracy. (Iqbal et al., 2020) introduced CSLNet, detecting complex lesion patterns with an AUROC of 99.1%, but omitted metadata, potentially limiting performance.

2.3.4. Explainability

Meia et al. used ABELE for explainability but highlighted the need for real-world validation (Metta et al., 2021). (Nigar et al., 2022) applied LIME with ResNet-18, achieving 94.47% accuracy but lacking healthy skin samples for contrast.

2.3.5. Hybrid Models

(Olayah et al., 2023) employed geometric active contour segmentation with hybrid CNN architectures, achieving 96.1% accuracy. (Villa-Pulgarin et al., 2021) used DenseNet-201 for classification with 93% accuracy but noted limited preprocessing as a constraint.

Overall, these studies highlight the importance of dataset size, metadata, explainability, and hybrid approaches in advancing skin lesion classification.

2.4. Literature Review Discussion

A detailed analysis of 22 studies reveals key trends in skin lesion classification research (Table 1-3). While traditional machine learning models like decision trees, SVM, and random forests were used in a few studies e.g. (Allugunti, 2022; Swamy & Divya, 2021), the majority employed deep learning

architectures, particularly trained CNNs. (Ahmad et al., 2023) achieved the highest accuracy of 99.3% on the HAM10000 dataset using a fusion model of Xception and ShuffleNet. In ISIC 2019 dataset studies, (Gong et al., 2020) attained the best accuracy of 99.50% with their DecisionFusion3 model. The ISIC 2019 challenge winner, (Gessert et al., 2020), excelled in metrics like AUC, sensitivity, and specificity, with a maximum AUC of 98.00%.

Feature extraction methods were explicitly noted in only four studies. Three studies employed color extraction except (Ahmad et al., 2020), which used CNN-calculated embeddings), while (Swamy & Divya, 2021) also included texture features. Most other works relied on automatic feature extraction by CNNs. ResNet emerged as the most popular CNN, appearing in seven studies, followed by DenseNet (4 studies) and Inception, Inception-ResNet, and GoogLeNet (3 studies each). Fusion models, particularly ensembles, showed superior performance. Three studies combined segmentation with classification (Hoang et al., 2022), reporting better outcomes compared to non-segmented approaches. For explainability, seven techniques were applied, with GradCAM being the most common, used in three papers. Saliency maps, attention modules, and SHAP were also noted, with Ahmad et al. (Ahmad et al., 2023) achieving the best accuracy among GradCAM-enabled models.

Despite advancements, limitations persist. Dataset imbalance, especially in the ISIC 2019 dataset where the melanoma nevus class comprises 73% of the data, and the lack of diverse skin tones and metadata are significant challenges. Only one study (Gessert et al., 2020) incorporated patient metadata like age and gender. Future work must prioritize diverse, metadata-rich datasets and explainable models.

2.5. Theoretical Background

2.5.1. MobileNetV3

MobileNetV3 (Howard et al., 2019) is optimized for mobile and edge devices, employing depthwise separable convolutions and neural architecture search (NAS) for efficiency. Features like squeeze-and-excitation modules and hard swish activation enhance performance while maintaining low computational cost.

2.5.2. Saliency Maps

Saliency maps (Simonyan et al., 2013) visualize input regions crucial for a model's decisions by computing gradients of the output score concerning input pixels. These maps, often grayscale, reveal influential areas and aid in understanding model focus and biases.

3. Critical Analysis Summary

Tables 1–3 provide a structured comparison of datasets, feature extraction methods, algorithms, explainability models, and performance metrics across the 22 studies. These insights guide improvements in skin lesion classification.

Table 1. Summary of related studies to skin lesion classification

Citations	Dataset	Feature representation	Algorithm	Results	Strength	Weakness
(Allugunti, et al., 2022)	Dermnet	-	Decision trees, Random Forest, Gradient Boosted Trees, CNN	Best results: CNN PRE: 91.07% REC: 87.86% F1: 89.32% ACC: 88.83%	High precision, recall, accuracy, and F1-scores The model looks promising for multiclass classification.	Unclear feature representation

(Ahmad et al., 2023)	AI-skin	Embeddings calculated by CNN and triplet loss function	ResNet152 + Triplet, Inception ResNet-V2 + Triplet	Best results: Inception ResNet-V2 + Triplet ACC: 87.42% REC: 97.04% SPE: 96.48%	The model outperforms SOTA works for skin disease classification	The model can be improved with a better dataset curated by dermatologist
(Swamy & Divya, 2021)	Dermnet, Dermquest	Color and texture feature extraction	Decision trees, SVM	Best results: SVM ACC: Color: 75% Texture: 83%	Texture feature extraction increased accuracy, and implemented simple machine learning models.	A larger image database is necessary for better results, and better feature extraction needed
(Wu et al., 2019)	Xiangya-Derm	-	ResNet50, Inception V3, DenseNet-121, Xception, Inception ResNet V2	Best results: Inception ResNet V2 PRE: 70.8% REC:77.0%	CNN architectures showed overall satisfactory results, pre-trained CNNs perform better than non	Precision and recall for AK class of best model low, datasets quality and quantity must be improved.

Table 2. Summary of related studies to explainable skin lesion classification

Citations	**Dataset**	**Feature representation**	**Algorithm**	**Explainability model**	**Results**	**Strength**	**Weakness**
(Ballari et al., 2022)	Skin disease dataset from Kaggle	-	ResNet-18	GradCA	Model accuracy of 96%, Grad-CAM output displays a convolutional feature map.	High model accuracy, visually interpretable results	Results not communicated by author, dataset not specified, unknown dataset size

(Barata et al., 2020)	ISIC 2017, ISIC 2018	Color normalization	Hierarchical taxonomy method, VGG-16, ResNet-50, DenseNet-161	Trainable attention	Best results on ISIC 2017: VGG-16 SEN: 86.7% , SPE: 87.1% ,AUC: 92.4%	Hierarchical taxonomy bred competitive results, the model correctly identifies relevant lesion regions, and color normalization has proven to improve accuracy.	Melanoma class not easily identified, model not robust to withstand varying transformations on images
(Ding et al., 2023)	ISIC-2017, ISIC-2018, HAM10000	-	HI-MViT	GradCAM, AblationCAM,	Best results on ISIC 2018 dataset PRE: 93.1% REC: 93.2% F1: 93.1% , ACC: 93.2%	With high results, the model can be generalized well.	The dataset lacks diversity in skin tones and does not have other medical indicators.
(Ahmad et al. 2023)	HAM10000, ISIC 2018	-	Xception, Shufflenet , a fusion of both models Features selected and fused using the IBOA method	GradCAM	Best results HAM10000 dataset, fusion model ACC: 99.3%, REC: 99.38% PRE: 99.4% , F1: 99.38%	High accuracy for HAM10000 dataset, GradCAM visualization clear	Bayesian optimization may improve results

(Young et al., 2019)	HAM10000	-	Inception	GradCAM, KernelSHAP	AUC (mean): 85%, REC (mean): 87%	High average AUC and recall across 30 different models	Small dataset size led to spurious correlations by model, Inceptiononly model limited in accuracy ceiling

Table 3: Summary of related studies to ISIC 2019 classification dataset

Citation	Dataset(s)	Feature Representation	Algorithm	Explainability Model	Results	Strengths	Weaknesses
(El-Khatib et al., 2020)	ISIC 2019, PH2	Histogram of Oriented Gradient	GoogLeNet, ResNet-101, NasNet-Large, decision fusion	-	ACC: 93.00%, SPE: 93.33%, SEN: 92.50%	The fusion model outperformed individual CNNs; with high accuracy, sensitivity, and specificity.	Small dataset (300 images total)
(Gessert et al., 2020)	ISIC 2019	-	EfficientNet	-	AUC: 98.00%, SEN: 55.60%, SPE: 99.30%	Metadata inclusion improved AUC and specificity; won the ISIC 2019 challenge.	Poor sensitivity; unreliable on out-of-distribution images
(Gong et al. 2020)	ISIC 2019, StyleGAN	-	Fusion CNN models	-	ACC: 99.50%, AUC: 98.90%, PRE: 98.40%, SEN: 98.30%, SPE: 99.60%	GANs addressed small datasets and class imbalance; with high fusion model accuracy.	The best fusion CNN combination is challenging to identify.
(Hoang et al. 2022)	HAM10000, ISIC 2019	-	EW-FCM segmentation, WideShuffleNet	-	ACC: 82.56%, SEN: 82.56%, SPE: 97.51%	Lightweight model with fewer parameters; better than non-segmented models	Underperformed compared to EfficientNet-B0

Iqbal et al.	ISIC 2017/2018/2019	-	CSLNet	-	ACC: 89.58%, AUROC: 99.10%, PRE: 90.66%, F1: 89.75%, SEN: 89.58%	Outperformed other models; multi-kernel design recognized symmetry and patterns	Lacks demographic factors like age, race, and gender
(Kassem et al., 2020)	ISIC 2019	-	GoogLeNet, GoogLeNet + SVM	-	ACC: 94.92%, SEN: 79.80%, SPE: 97.00%, PRE: 80.36%	Outperformed ISIC 2019 winning model	Multiclass SVM yielded lower test set performance
(Meia et al.,)	ISIC 2019	-	ResNet	ABELE	ACC: 83.80% (balanced multiclass)	Saliency maps enhance explainability	Lack of benchmarks for performance comparison
(Nigar et al., 2022)	ISIC 2019	-	ResNet-18	LIME	ACC: 94.47%, F1: 94.45%, PRE: 93.57%, REC: 94.01%	The model generalizes well; LIME provides visual explainability	Limited to 8 disease classes; lacks opposing examples
(Olayah et al., 2023)	ISIC 2019	-	GAC segmentation + hybrid CNN-ANN	-	ACC: 96.10%, AUC: 94.41%, SEN: 88.90%, SPE: 99.44%, PRE: 88.69%	Optimized segmentation and hybrid model with high accuracy	-
(Tô et al., 2019)	ISIC 2019	-	U-Net segmentation + EfficientNet-B4	-	-	-	-

(Villa-Pulgarin et al., 2021)	ISIC 2019, HAM10000	-	DenseNet-201, Inception-V3, InceptionResNet-V2	-	ACC: 93.00%, F1: 93.00%, PRE: 93.00%, REC: 93.00%	DenseNet-201 model comparable to state-of-the-art methods	

4. Comparative Analysis

The reviewed studies revealed significant variability in performance and methodological approaches. DenseNet-201 and fusion models incorporating metadata consistently outperformed image-only models, with accuracy rates exceeding 90%. Explainability techniques such as Grad-CAM were frequently employed, offering visual clarity but occasionally at the cost of reduced predictive power. Gong et al. (2020) demonstrated that augmenting datasets with synthetic images from StyleGAN could mitigate class imbalance, significantly boosting accuracy to 99.5%. Models like MobileNetV3, despite their computational efficiency, struggled with generalizability due to limited dataset diversity. Studies leveraging hybrid architectures and advanced optimization techniques achieved superior results, underscoring the importance of combining robust model designs with diverse, high-quality datasets. These comparative insights emphasize the need for balanced methodologies that prioritize both accuracy and interpretability (Figure 2).

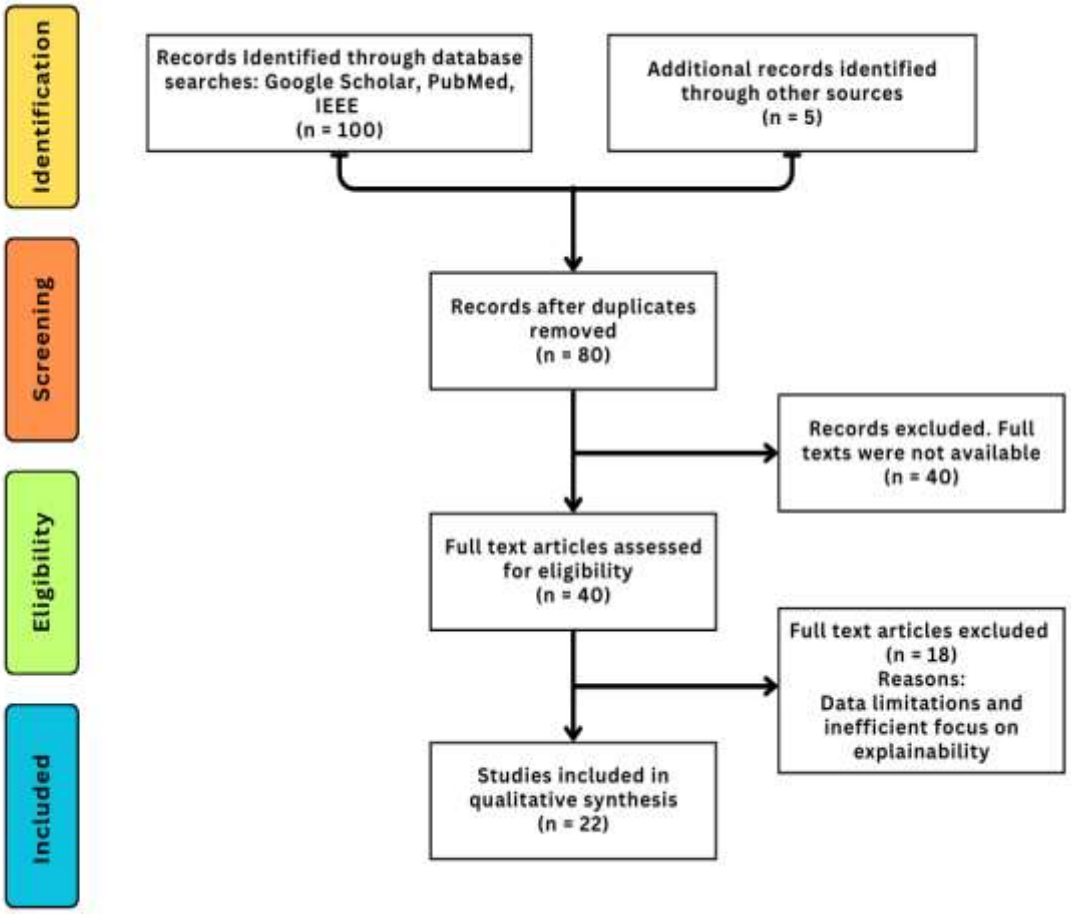

Fig 1. Illustrating PRISMA flow diagram for the systematic selection of studies, from initial screening to final inclusion.

5. Results and Discussions

5.1. Results

The study evaluated 22 research papers focusing on integrating XAI into skin lesion classification. The most common explainability methods included Grad-CAM, SHAP, and saliency maps, which provided visual insights into model decisions. DenseNet-201 models achieved a high accuracy of 93% when combined with metadata, highlighting the value of incorporating additional contextual information. Gong et al. (2020) utilized StyleGAN-generated synthetic data to address class imbalances, achieving a record accuracy of 99.5%. Conversely, the MLP-MobileNetV3 implementation fell short of baseline performance, with only 40% accuracy and a mean sensitivity of 54%, largely due to dataset limitations and insufficient model complexity. These findings underline the trade-offs between accuracy and interpretability in XAI applications, suggesting that advanced optimization techniques and diverse datasets are critical for improving outcomes.

5.2. Discussion and Practical Implications

The integration of XAI in skin lesion classification has demonstrated both potential and challenges. Explainability methods, such as Grad-CAM and SHAP, enhance transparency, enabling clinicians to validate model decisions. However, achieving high interpretability often results in compromises in accuracy, highlighting the need for balanced solutions. Practical hurdles, including dataset imbalances and lack of diversity in metadata, limit model generalizability across populations. Additionally, regulatory requirements for clinical AI tools necessitate rigorous validation and standardization. Future research should focus on interdisciplinary collaboration to bridge technical advancements and clinical needs. Prioritizing diverse datasets, leveraging synthetic data, and refining hybrid architectures will facilitate the development of reliable, interpretable AI-driven diagnostic tools. To advance the field, the following steps are recommended: Employ optimization methods like Bayesian techniques to improve model efficiency and accuracy. Incorporate synthetic data generation to address dataset imbalances and enhance diversity. Develop hybrid architectures that combine metadata integration with advanced augmentation strategies. Establish standardized benchmarks for evaluating XAI models in clinical applications.Foster interdisciplinary research collaborations to align AI innovations with clinical requirements.

6. Conclusion

Explainable AI (XAI) represents a pivotal advancement in the field of skin lesion classification, addressing the critical need for transparency and trust in AI-driven healthcare. This review highlights the significant strides made by integrating XAI into machine learning models, with techniques such as Grad-CAM and SHAP enhancing interpretability while providing actionable insights for clinicians. However, the findings also reveal persistent challenges, including dataset imbalances, limited diversity in metadata, and the trade-offs between interpretability and predictive accuracy. The reviewed studies emphasize the importance of diverse, high-quality datasets and advanced optimization methods in achieving robust performance. Models leveraging metadata and synthetic data have demonstrated superior accuracy and generalizability, yet their implementation in clinical settings remains constrained by regulatory and practical limitations. Addressing these issues requires a multidisciplinary approach, combining expertise from computer science, medicine, and regulatory affairs.

Future research should focus on developing hybrid architectures that balance accuracy and interpretability, leveraging innovative data augmentation techniques, and fostering standardization in model evaluation. By aligning technical advancements with clinical requirements, XAI can transform dermatological diagnostics, making AI tools not only more effective but also more acceptable to healthcare practitioners and patients. XAI's potential to bridge the gap between complex machine learning models and clinical applicability is undeniable. While challenges remain, the path forward involves prioritizing data diversity, methodological rigor, and interdisciplinary collaboration. These efforts will ensure that XAI becomes an integral part of trustworthy, efficient, and interpretable AI-driven healthcare solutions, ultimately improving patient outcomes on a global scale.

References

1. Ahmad, B., Usama, M., Huang, C., Hwang, K., Hossain, M. S., & Muhammad, G. (2020). Discriminative feature learning for skin disease classification using deep convolutional neural network. IEEE Access, 8, 39025–39033. https://doi.org/10.1109/access.2020.2975198
2. Ahmad, N., Shah, J. H., Khan, M. A., Baili, J., Ansari, G. J., Tariq, U., Kim, Y. J., & Cha, J. (2023). A novel framework of multiclass skin lesion recognition from dermoscopic images using deep learning and explainable AI. Frontiers in Oncology, 13. https://doi.org/10.3389/fonc.2023.1151257
3. Allugunti, V. R. (2022). A machine learning model for skin disease classification using convolution neural network. International Journal of Computing Programming and Database Management, 3(1), 141–147. https://doi.org/10.33545/27076636.2022.v3.i1b.53
4. Ballari, G. S., Giraddi, S., Chickerur, S., & Kanakareddi, S. (2022). An explainable AI-based skin disease detection. In Lecture Notes in Networks and Systems (pp. 287–295). Springer. https://doi.org/10.1007/978-981-19-5331-6_30
5. Barata, C., Celebi, M. E., & Marques, J. S. (2020). Explainable skin lesion diagnosis using taxonomies. Pattern Recognition, 110, 107413. https://doi.org/10.1016/j.patcog.2020.107413
6. Codella, N. C. F., Rotemberg, V., Tschandl, P., Celebi, M. E., Dusza, S. W., Gutman, D. A., Helba, B., Kalloo, A., Liopyris, K., Marchetti, M. A., Kittler, H., & Halpern, A. (2019). Skin lesion analysis toward melanoma detection 2018: A challenge hosted by the International Skin Imaging Collaboration (ISIC). arXiv. https://doi.org/10.48550/arxiv.1902.03368
7. Ding, Y., Yi, Z., Li, M., Long, J., Lei, S., Guo, Y., Fan, P., Zuo, C., & Wang, Y. (2023). HI-MViT: A lightweight model for explainable skin disease classification based on modified MobileViT. Digital Health, 9. https://doi.org/10.1177/20552076231207197
8. El-Khatib, H., Popescu, D., & Ichim, L. (2020). Deep learning–based methods for automatic diagnosis of skin lesions. Sensors, 20(6), 1753. https://doi.org/10.3390/s20061753
9. Gessert, N., Nielsen, M., Shaikh, M., Werner, R., & Schlaefer, A. (2020). Skin lesion classification using ensembles of multi-resolution EfficientNets with meta data. MethodsX, 7, 100864. https://doi.org/10.1016/j.mex.2020.100864
10. Gong, A., Yao, X., & Lin, W. (2020). Dermoscopy image classification based on StyleGANs and decision fusion. IEEE Access, 8, 70640–70650. https://doi.org/10.1109/access.2020.2986916
11. Hernández-Pérez, C., Combalia, M., Podlipnik, S., Codella, N. C. F., Rotemberg, V., Halpern, A. C., Reiter, O., Carrera, C., Barreiro, A., Helba, B., Puig, S., Vilaplana, V., & Malvehy, J. (2024). BCN20000: Dermoscopic lesions in the wild. Scientific Data, 11(1). https://doi.org/10.1038/s41597-024-03387-w
12. Hoang, L., Lee, S., Lee, E., & Kwon, K. (2022). Multiclass skin lesion classification using a novel lightweight deep learning framework for smart healthcare. Applied Sciences, 12(5), 2677. https://doi.org/10.3390/app12052677
13. Howard, A., Sandler, M., Chen, B., Wang, W., Chen, L., Tan, M., Chu, G., Vasudevan, V., Zhu, Y., Pang, R., Adam, H., & Le, Q. (2019). Searching for MobileNetV3. 2021 IEEE/CVF International Conference on Computer Vision (ICCV), 1314–1324. https://doi.org/10.1109/iccv.2019.00140
14. Iqbal, I., Younus, M., Walayat, K., Kakar, M. U., & Ma, J. (2020). Automated multi-class classification of skin lesions through deep convolutional neural network with dermoscopic images. Computerized Medical Imaging and Graphics, 88, 101843. https://doi.org/10.1016/j.compmedimag.2020.101843
15. Kassem, M. A., Hosny, K. M., & Fouad, M. M. (2020). Skin lesions classification into eight classes for ISIC 2019 using deep convolutional neural network and transfer learning. IEEE Access, 8, 114822–114832. https://doi.org/10.1109/access.2020.3003890
16. Kurasinski, L., & Mihailescu, R. (2020). Towards machine learning explainability in text classification for fake news detection. 2021 20th IEEE International Conference on Machine Learning and Applications (ICMLA). https://doi.org/10.1109/icmla51294.2020.00127
17. Lu, Y., & Li, C. (2020). GCAN: Graph-aware co-attention networks for explainable fake news detection on social media. arXiv. https://doi.org/10.48550/arxiv.2004.11648

18. Metta, C., Beretta, A., Guidotti, R., Yin, Y., Gallinari, P., Rinzivillo, S., & Giannotti, F. (2021). Explainable deep image classifiers for skin lesion diagnosis. arXiv. https://doi.org/10.48550/arxiv.2111.11863
19. Ni, S., Li, J., & Kao, H. (2021). MVAN: Multi-view attention networks for fake news detection on social media. IEEE Access, 9, 106907–106917. https://doi.org/10.1109/access.2021.3100245
20. Nigar, N., Umar, M., Shahzad, M. K., Islam, S., & Abalo, D. (2022). A deep learning approach based on explainable artificial intelligence for skin lesion classification. IEEE Access, 10, 113715–113727. https://doi.org/10.1109/access.2022.3217217
21. Olayah, F., Senan, E. M., Ahmed, I. A., & Awaji, B. (2023). AI techniques of dermoscopy image analysis for the early detection of skin lesions based on combined CNN features. Diagnostics, 13(7), 1314. https://doi.org/10.3390/diagnostics13071314
22. Rehman, M. Z. U., Ahmed, F., Alsuhibany, S. A., Jamal, S. S., Ali, M. Z., & Ahmad, J. (2022). Classification of skin cancer lesions using explainable deep learning. Sensors, 22(18), 6915. https://doi.org/10.3390/s22186915
23. Simonyan, K., Vedaldi, A., & Zisserman, A. (2013). Deep inside convolutional networks: Visualising image classification models and saliency maps. arXiv. https://doi.org/10.48550/arxiv.1312.6034
24. Swamy, K. V., & Divya, B. (2021, December 16). Skin disease classification using machine learning algorithms. In ISIC2019. https://doi.org/10.1109/c2i454156.2021.9689338
25. Tô, T. D., Lan, D. T., Nguyen, T. T. H., Nguyen, T. T. N., Nguyen, H., Phuong, L. B., & Nguyen, T. Z. (2019, October 28). Ensembled skin cancer classification (ISIC 2019 challenge submission). HAL Archives. https://hal.science/hal-02335240v1
26. Tschandl, P., Argenziano, G., Razmara, M., & Yap, J. (2018). Diagnostic accuracy of content-based dermatoscopic image retrieval with deep classification features. British Journal of Dermatology, 181(1), 155–165. https://doi.org/10.1111/bjd.17189
27. Tschandl, P., Rosendahl, C., & Kittler, H. (2018). The HAM10000 dataset, a large collection of multi-source dermatoscopic images of common pigmented skin lesions. Scientific Data, 5(1), Article 161. https://doi.org/10.1038/sdata.2018.161
28. Vavekanand, R., & Kumar, T. (2024). Data augmentation of ultrasound imaging for non-invasive white blood cell in vitro peritoneal dialysis. Biomedical Engineering Communications, 3(4), Article 17. https://doi.org/10.53388/bmec2024017
29. Vavekanand, R., Sam, K., Kumar, S., & Kumar, T. (2024). CardiacNet: A neural networks-based heartbeat classification using ECG signals. Studies in Medical and Health Sciences, 1(2), 1–17. https://doi.org/10.48185/smhs.v1i2.1188
30. Villa-Pulgarin, J. P., Ruales-Torres, A. A., Arias-Garz, D., Bravo-Ortiz, M. A., Arteaga-Arteaga, H. B., Mora-Rubio, A., Alzate-Grisales, J. A., Mercado-Ruiz, E., Hassaballah, M., Orozco-Arias, S., Cardona-Morales, O., & Tabares-Soto, R. (2021). Optimized convolutional neural network models for skin lesion classification. Computers, Materials & Continua, 70(2), 2131–2148. https://doi.org/10.32604/cmc.2022.019529
31. Wu, Z., Zhao, S., Peng, Y., He, X., Zhao, X., Huang, K., Wu, X., Fan, W., Li, F., Chen, M., Li, J., Huang, W., Chen, X., & Li, Y. (2019). Studies on different CNN algorithms for face skin disease classification based on clinical images. IEEE Access, 7, 66505–66511. https://doi.org/10.1109/access.2019.2918221
32. Yang, F., Pentyala, S. K., Mohseni, S., Du, M., Yuan, H., Linder, R., Ragan, E. D., Ji, S., & Hu, X. (2019, May 13). XFake: Explainable fake news detector with visualizations. Proceedings of the World Wide Web Conference (WWW '19). https://doi.org/10.1145/3308558.3314119
33. Young, K., Booth, G., Simpson, B., Dutton, R., & Shrapnel, S. (2019). Deep neural network or dermatologist? In Lecture Notes in Computer Science (pp. 48–55). https://doi.org/10.1007/978-3-030-33850-3_6

34. Barata, C., Marques, J. S., & Celebi, M. E. (2019). Deep attention model for the hierarchical diagnosis of skin lesions. In 2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW) (pp. 2757–2765). https://doi.org/10.1109/cvprw.2019.00334
35. Vavekanand, R. (2024). A Machine Learning Approach for Imputing ECG Missing Healthcare Data. SSRN Electronic Journal. https://doi.org/10.2139/ssrn.4822530
36. Singh, V., Anwar, S., & Vavekanand, R. (2024). Recognizing Mitral Regurgitation Through Machine Learning Algorithms in Cardiac Imaging. Research Square (Research Square). https://doi.org/10.21203/rs.3.rs-4586175/v1
37. Vavekanand, R. (2024). SUBMIP: Smart Human Body Health Prediction Application System Based on Medical Image Processing. Studies in Medical and Health Sciences, 1(1), 14–22. https://doi.org/10.48185/smhs.v1i1.1141