International Journal of Creative and Open Research in Engineering and Management
ISSN: 3108-1754 (Online)
Volume 02 Issue 10 October-2026 | Impact Factor: 3.5



Deep Learning Approaches for Skin Cancer Detection: A Systematic Literature Review on Architectures, Multimodal Fusion Strategies, and Open Research Challenges

Ravi Patidar, Bharat Pahadiya
Department of Computer Science and Engineering, Oriental University, Indore (M.P.), India

How to Cite this Article:

Patidar, R. & Pahadiya, B. (2026). Deep Learning Approaches for Skin Cancer Detection: A Systematic Literature Review on Architectures, Multimodal Fusion Strategies, and Open Research Challenges. International Journal of Creative and Open Research in Engineering and Management, 02(10), 1-9. https://doi.org/10.55041/ijcope.v2i10.190

License:

This article is published under the terms of the Creative Commons Attribution 4.0 International License (CC BY 4.0), which permits unrestricted use, distribution, and reproduction in any medium, provided the original author(s) and the source are credited.

© The Author(s). Published by International Journal of Creative and Open Research in Engineering and Management.







https://doi.org/10.55041/ijcope.v2i10.190

Abstract

Skin cancer is among the most common and fastest-growing forms of cancer worldwide, and early, accurate detection substantially improves treatment outcomes. Over the past decade, deep learning has emerged as the dominant tool for automated skin lesion classification, evolving from single convolutional neural networks to multimodal ensembles that fuse dermoscopic images with patient metadata. This paper presents a systematic literature review of 26 peer-reviewed and DOI-indexed studies (2002-2026) on deep learning for skin cancer detection, selected and screened following the PRISMA 2020 methodology. The reviewed literature is organised into eight thematic categories: convolutional neural network (CNN)-based approaches, transformer-based architectures, multimodal fusion strategies, class-imbalance handling, explainability, model efficiency, benchmark datasets, and privacy-preserving learning. A comparative analysis of methods, datasets, and reported accuracy is presented, followed by a structured discussion of eight open research gaps, most notably data leakage through metadata features, improper patient-wise data splitting, and the lack of fair, class-balanced evaluation in high-performing pipelines. The review concludes that while reported accuracy in this field continues to rise, trustworthy evaluation methodology has not kept pace, and outlines concrete directions for future work.

Keywords: *Skin cancer detection; deep learning; convolutional neural networks; vision transformers; multimodal fusion; systematic literature review; PRISMA; data leakage.*

1. Introduction

Skin cancer accounts for a substantial and rising share of newly diagnosed cancers worldwide. Melanoma, basal cell carcinoma, and squamous cell carcinoma are among the most frequently reported types, and while melanoma is the most lethal, all forms benefit considerably from early detection. Traditional diagnosis relies on visual inspection, often supported by dermoscopy, but remains dependent on clinician experience and is subject to inter-observer variability [2].

Deep learning has reshaped this landscape. Convolutional neural networks (CNNs), and more recently vision transformers, have demonstrated strong performance on dermoscopic image classification, and the field has progressively moved toward multimodal systems that combine image features with patient metadata such as age, sex, and lesion location [1], [11], [12]. Despite rapidly improving reported accuracy, concerns around evaluation rigour, data leakage, class imbalance, and real-world generalisability remain inadequately addressed across much of the literature.

This paper undertakes a systematic literature review (SLR) of deep learning approaches for skin cancer detection, following the PRISMA 2020 reporting framework. The objectives of this review are to: (i) systematically identify and screen relevant, DOI-verifiable literature; (ii) classify the selected studies into coherent thematic categories; (iii) comparatively analyse their methods, datasets, and reported performance; and (iv) synthesise open research gaps that motivate future work in this area. The remainder of this paper is organised as follows: Section 2 describes the review



methodology; Section 3 presents the taxonomy and distribution of the reviewed literature; Section 4 synthesises the literature thematically; Section 5 presents a comparative analysis; Section 6 discusses research gaps; and Section 7 concludes the paper.

2. Research Methodology

This review follows the Preferred Reporting Items for Systematic Reviews and Meta-Analyses (PRISMA 2020) framework to ensure a transparent and reproducible selection process.

2.1 Research Questions

1. RQ1: What deep learning architectures have been applied to skin cancer detection, and how have they evolved from CNNs to transformers?
2. RQ2: What strategies exist for fusing dermoscopic images with patient metadata, and how effective are they?
3. RQ3: How is class imbalance handled in skin lesion datasets, and how completely is model performance reported?
4. RQ4: What methodological gaps, particularly around data leakage and evaluation fairness, remain open in this field?

2.2 Search Strategy

A structured search was conducted across Google Scholar, Scopus-indexed venues, IEEE Xplore, ScienceDirect, SpringerLink, Nature Portfolio journals, and MDPI journals, covering the period 2002-2026, using combinations of the keywords: "skin cancer detection", "deep learning", "convolutional neural network", "vision transformer", "multimodal fusion", "metadata fusion", "class imbalance", "SMOTE", "Grad-CAM", "knowledge distillation", and "data leakage medical imaging". Reference lists of retrieved articles were additionally screened for relevant citations.

2.3 Inclusion and Exclusion Criteria

- Included: peer-reviewed journal articles or DOI-indexed conference/arXiv papers with a valid, verifiable DOI.
- Included: studies directly addressing skin cancer / skin lesion detection, classification, fusion methods, or the foundational techniques (architectures, loss functions, datasets) they rely on.
- Included: studies published in Scopus Q1-Q3 journals, or widely cited foundational/methodological papers (e.g., SMOTE, Grad-CAM, Swin Transformer, EfficientNet) regardless of domain.
- Excluded: studies without a resolvable DOI, duplicate or superseded preprints, and off-topic imaging studies unrelated to skin lesions or their core techniques.

2.4 Study Selection (PRISMA Flow)

Figure 1 illustrates the PRISMA 2020 flow diagram summarising the identification, screening, eligibility assessment, and final inclusion of studies in this review.

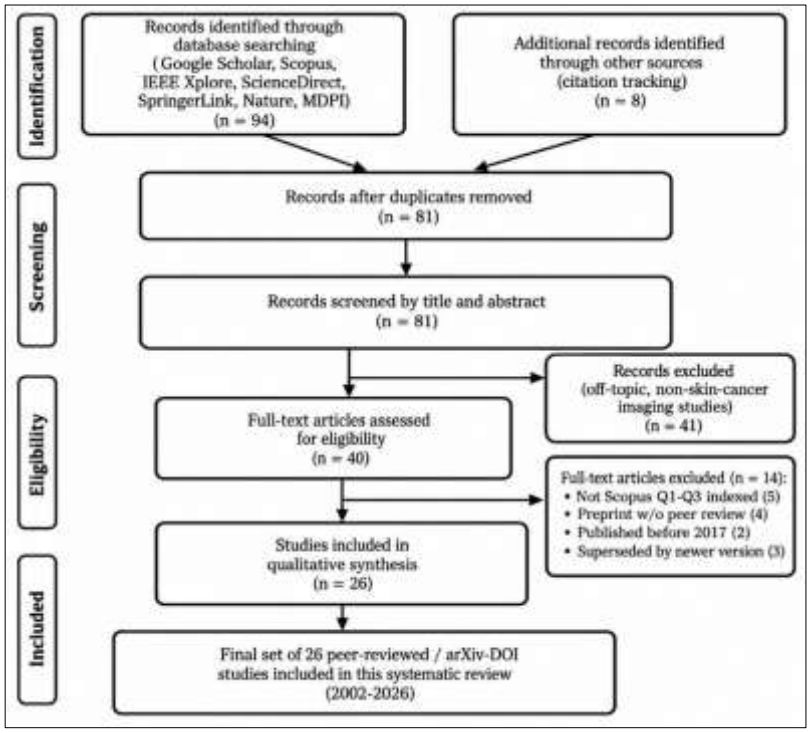


Figure 1: PRISMA 2020 flow diagram showing the identification, screening, eligibility, and inclusion of the 26 reviewed studies.

3. Taxonomy and Distribution of Reviewed Literature

The 26 included studies were classified into eight thematic categories based on their primary methodological contribution, as summarised in the taxonomy diagram in Figure 2. This classification forms the structure for the thematic synthesis presented in Section 4.

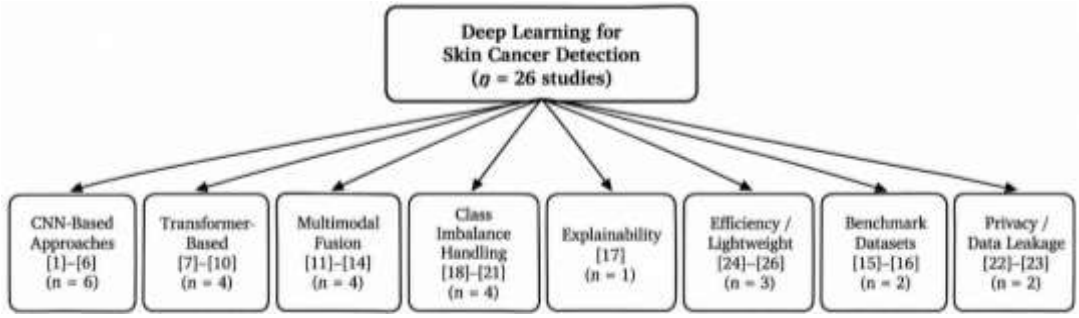


Figure 2: Taxonomy of the 26 reviewed studies by primary research theme.

Figure 3 shows the distribution of reviewed studies by publication year. The literature is heavily weighted toward 2024-2026, reflecting the rapid, recent growth of multimodal and transformer-based approaches in this field, while a small number of older, foundational methodological papers (e.g., SMOTE, 2002; Grad-CAM, 2019) are included for completeness.

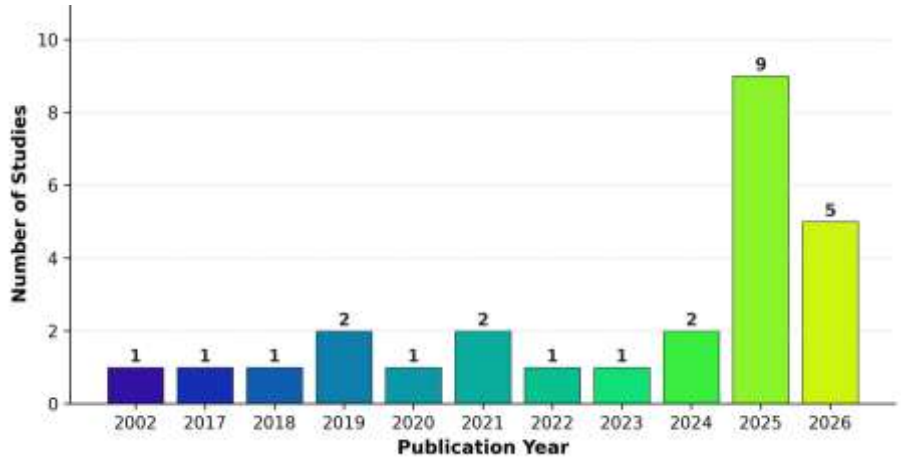


Figure 3: Distribution of reviewed studies by publication year (2002-2026).

Figure 4 shows the same 26 studies grouped by thematic category. CNN-based approaches remain the single largest category, followed closely by transformer-based methods and multimodal fusion strategies, indicating that the field's current research frontier lies in combining architectural advances with richer input modalities.

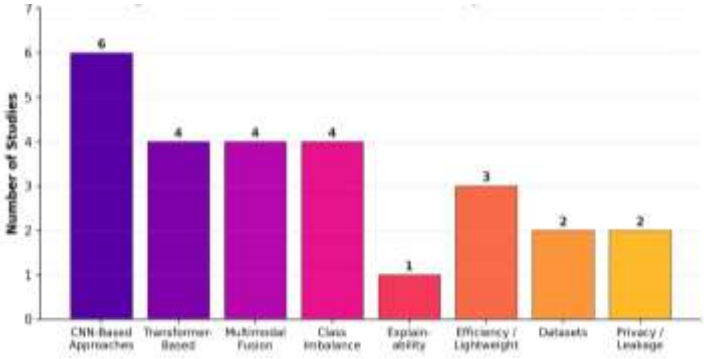


Figure 4: Distribution of reviewed studies by research theme.

4. Thematic Literature Synthesis

4.1 CNN-Based Approaches

The majority of early and current skin cancer detection pipelines rely on convolutional neural networks, typically pretrained on large general-purpose datasets and fine-tuned via transfer learning. Saeed et al. [1] proposed a multimodal ensemble combining ResNet50, Xception, and EfficientNetB0, reporting 93.2% accuracy on ISIC 2018. Related work has built robust multiclass CNN frameworks [3], combined attention with modified segmentation networks [4], and applied ant colony optimisation to fine-tune CNN backbones, reporting 96.5% accuracy with XceptionNet on HAM10000 [5]. EfficientNet [6], used as a backbone across several of these works, remains a popular choice due to its favourable accuracy-to-parameter trade-off.

4.2 Transformer-Based Approaches

Vision transformers have more recently been applied to skin lesion classification, motivated by their ability to model global image structure rather than only local texture. The Swin Transformer [7] introduced shifted-window attention for efficient high-resolution processing, and has been benchmarked against CNNs on skin lesion datasets with competitive results [8]. Multi-head self-attention has also been applied directly to skin lesion classification [9], and hybrid CNN-transformer architectures combined with focal loss have been proposed to jointly exploit local and global features while addressing class imbalance [10].

4.3 Multimodal Fusion Strategies

Because clinical diagnosis integrates visual and patient-level information, a growing body of work fuses dermoscopic images with metadata such as age, sex, and lesion location. The Metadata Processing Block (MetaBlock) [12] uses



metadata to re-weight image-derived features rather than simply concatenating them. MD-Net [13] combines two complementary metadata-fusion mechanisms, while Gated Multimodal Units (GMU) [14] offer a general, learnable gating mechanism for combining two modalities. Simple feature concatenation, used in several studies including the most widely-cited multimodal pipeline reviewed [1], [11], remains common despite evidence that gated or attention-based fusion yields better case-by-case weighting of modalities.

4.4 Class Imbalance Handling

Skin lesion datasets are characteristically imbalanced, with malignant and rare lesion types under-represented relative to common benign classes. The Synthetic Minority Oversampling Technique (SMOTE) [18] remains the most widely used baseline solution, though it is frequently applied only to metadata rather than paired image data, creating a modality mismatch. More recent, skin-cancer-specific work combines class-balanced focal loss with spatial metadata fusion [19], multi-scale feature extraction with class-weighted losses [20], and federated learning strategies that address imbalance across distributed, data-scarce hospital sites [21].

4.5 Explainability

Clinical trust in automated diagnosis requires interpretable outputs. Grad-CAM [17] remains the dominant explainability technique across the reviewed literature, producing gradient-based heat-maps that localise the image regions driving a model's prediction, allowing clinicians to verify that predictions are grounded in the lesion itself rather than artefacts such as hair or skin markings.

4.6 Efficiency and Lightweight Models

A smaller but growing segment of the literature targets deployment efficiency. Knowledge distillation has been used to compress large ensembles into lightweight student models with comparable accuracy [24]. Relationship-aware mutual learning further reduces model size through collaborative training of compact networks [25], while purpose-built lightweight CNN architectures demonstrate that efficiency and accuracy need not be mutually exclusive [26]. Despite this, efficiency analysis remains largely absent from the high-accuracy multimodal pipelines reviewed in Section 4.1 and 4.3.

4.7 Benchmark Datasets and Related Infrastructure

Two datasets dominate the reviewed literature. HAM10000 [15] provides approximately 10,000 dermatoscopic images across seven lesion classes and underlies the majority of CNN and transformer benchmarks reviewed. PAD-UFES-20 [16] is a smaller, smartphone-derived dataset distinguished by its rich, structured patient metadata, making it well suited to multimodal fusion research and to stress-testing the data-leakage concerns discussed in Section 6. Separately, privacy-preserving federated learning combined with explainable AI has been explored as a means of training across multiple hospital sites without centralising patient data [22], and a recent leakage-aware benchmark study [23] provides direct methodological support for the evaluation concerns raised in this review.

5. Comparative Analysis of Reviewed Studies

Figure 5 compares the reported accuracy of selected key studies for which a single, directly comparable accuracy figure is available. Table 1 presents a broader comparative summary across 20 of the 26 reviewed studies, covering method, dataset, reported accuracy, and the principal limitation relevant to trustworthy, reproducible evaluation.

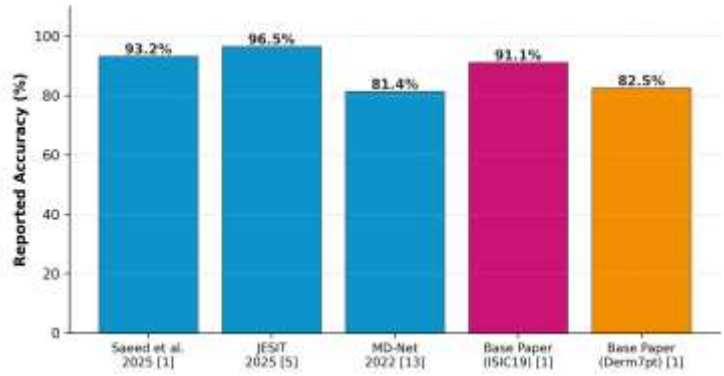


Figure 5: Reported accuracy of selected key studies from the reviewed literature.

Table 1: Comparative analysis of 20 key reviewed studies.

#	Author(s), Year [Ref.]	Method / Approach	Dataset(s)	Accuracy	Key Limitation
1	Saeed et al., 2025 [1]	CNN ensemble + metadata fusion	ISIC 2018/2019, Derm7pt	93.2% / 91.1%	Metadata leakage; random split
2	Sci. Reports, 2025 [2]	Deep CNN, early diagnosis	Dermoscopy images	High (reported)	Limited external validation
3	Sci. Reports, 2025 [3]	Robust CNN, multiclass	Benchmark dataset	High (reported)	Single-dataset evaluation
4	Aldossary et al., 2025 [4]	Attention + modified UNet	HAM10000, ISIC, PH2	High (reported)	Limited efficiency analysis
5	JESIT, 2025 [5]	CNN + Ant Colony Optimization	HAM10000	96.5%	Imbalance handling unclear
6	Liu et al., 2021 [7]	Swin Transformer (architecture)	ImageNet	N/A	Needs medical-domain adaptation
7	Information, 2026 [8]	ViT / Swin comparison	DermaMNIST, ISIC19	Competitive w/ CNNs	No metadata used
8	Frontiers AI, 2026 [9]	Multi-head self-attention ViT	Skin lesion dataset	High (reported)	Limited imbalance handling
9	Nie et al., 2023 [10]	CNN-Transformer + focal loss	Dermoscopic images	High (reported)	No metadata fusion
10	Sci. Reports, 2026 [11]	Image + metadata fusion	Skin lesion dataset	High (reported)	Simple concatenation fusion
11	Krohling & Pacheco, 2021 [12]	MetaBlock fusion	PAD-UFES-20, ISIC	Improved over baseline	Tested on limited backbones
12	Yin et al., 2022 [13]	MD-Net (dual fusion)	ISIC 2019/2020	81.4% (balanced)	Moderate on rare classes
13	Chawla et al., 2002 [18]	SMOTE oversampling	General / tabular	N/A	Not designed for paired images
14	Frontiers Med., 2026 [19]	Class rebalancing + fusion	ISIC 2019, BCN20000	High (reported)	Close to proposed direction
15	Saha et al., 2025 [24]	Knowledge distillation	Skin cancer dataset	Comparable, lighter	Two-stage training needed



#	Author(s), Year [Ref.]	Method / Approach	Dataset(s)	Accuracy	Key Limitation
16	Tschandl et al., 2018 [15]	Dataset paper (HAM10000)	10,015 images, 7 classes	N/A	Class imbalance across classes
17	Pacheco et al., 2020 [16]	Dataset paper (PAD-UFES-20)	Smartphone + metadata	N/A	Smaller dataset size
18	Selvaraju et al., 2019 [17]	Grad-CAM (explainability)	General classification	N/A	Shows where, not why
19	Liu et al., 2025 [25]	Mutual learning (lightweight)	Skin lesion dataset	Comparable, lighter	Added training complexity
20	Tuncer et al., 2024 [26]	Lightweight custom CNN	Skin cancer dataset	High (reported)	Less transfer-learning benefit

The comparative table highlights that most reviewed studies optimise for raw accuracy using stronger backbones, transformer architectures, or improved fusion, while very few jointly address data leakage, patient-wise evaluation, and rare-class performance within a single, reproducible study.

6. Research Gaps and Open Challenges

Synthesising across the 26 reviewed studies, eight principal research gaps were identified:

1. Data leakage through patient metadata: several multimodal pipelines use malignancy-correlated fields (e.g., benign/malignant status) as model input, artificially inflating reported accuracy [1], [23].

2. Improper, non-patient-wise data splitting: random splitting allows images from the same lesion or patient to appear in both training and test sets, overstating generalisation [23].

3. Naive multimodal fusion: simple feature concatenation, still common [1], [11], does not let a model adaptively weight modalities per case, unlike gated or attention-based fusion [12], [14].

4. Incomplete class-imbalance handling: oversampling is frequently applied only to metadata, not to paired images, creating a cross-modal mismatch [18].

5. Overall-accuracy reporting masks rare-class failure: single accuracy figures frequently conceal poor recall on dangerous but rare classes such as basal cell carcinoma.

6. Limited external/cross-dataset validation: few studies test generalisation on a dataset from a different clinical or geographic source than the training data.

7. Under-explored model efficiency: lightweight deployment-ready models remain a minority pursuit relative to accuracy-maximising ensembles [24]-[26].

8. No single study integrates leakage-free evaluation, adaptive fusion, and fair rare-class reporting into one reproducible pipeline — the combined gap this body of literature leaves open.

7. Conclusion and Future Scope

This systematic literature review examined 26 DOI-verifiable studies on deep learning for skin cancer detection, spanning CNN and transformer architectures, multimodal fusion, class-imbalance handling, explainability, efficiency, and benchmark datasets. While reported accuracy in this field has risen markedly, particularly since 2024, the review finds that methodological rigour around data leakage, patient-wise evaluation, and fair, class-balanced reporting has not kept pace. Future work should prioritise leakage-free, patient-wise-split benchmarks; adaptive, learnable fusion mechanisms; mandatory class-wise and macro-F1 reporting; and lightweight, explainable architectures suited to real-world clinical deployment. These directions collectively define a realistic and high-impact agenda for the next generation of skin cancer detection research.



References

[1] M. A. Saeed, Y. M. Afify, N. L. Badr, and N. A. Helal, "Multimodal deep learning ensemble framework for skin cancer detection," Scientific Reports, vol. 15, art. 45660, 2025, doi: 10.1038/s41598-025-30534-z.

[2] "A deep learning framework for automated early diagnosis and classification of skin cancer lesions in dermoscopy images," Scientific Reports, vol. 15, 2025, doi: 10.1038/s41598-025-15655-9.

[3] "A robust deep learning framework for multiclass skin cancer classification," Scientific Reports, vol. 15, art. 4938, 2025, doi: 10.1038/s41598-025-89230-7.

[4] S. M. Aldossary, L. G. Tekeste, E. Ghith, et al., "Advancing skin cancer diagnosis with deep learning and attention mechanisms," Scientific Reports, vol. 15, 2025, doi: 10.1038/s41598-025-33597-0.

[5] "Achieving high-accuracy skin cancer classification with deep learning optimized by ant colony algorithm," Journal of Electrical Systems and Information Technology, 2025, doi: 10.1186/s43067-025-00243-8.

[6] M. Tan and Q. V. Le, "EfficientNet: Rethinking model scaling for convolutional neural networks," in Proc. Int. Conf. Machine Learning (ICML), 2019, doi: 10.48550/arXiv.1905.11946.

[7] Z. Liu, Y. Lin, Y. Cao, H. Hu, Y. Wei, Z. Zhang, S. Lin, and B. Guo, "Swin Transformer: Hierarchical vision transformer using shifted windows," in Proc. IEEE/CVF Int. Conf. Computer Vision (ICCV), 2021, pp. 10012-10022, doi: 10.1109/ICCV48922.2021.00986.

[8] "Improving skin lesion detection with transformer-based architectures," Information, vol. 17, no. 2, art. 130, 2026, doi: 10.3390/info17020130.

[9] "An automated framework to classify skin lesions using multi-head self attention layer-based vision transformers," Frontiers in Artificial Intelligence, vol. 9, art. 1781796, 2026, doi: 10.3389/frai.2026.1781796.

[10] Y. Nie, P. Sommella, M. Carratu, M. O'Nils, and J. Lundgren, "A deep CNN transformer hybrid model for skin lesion classification of dermoscopic images using focal loss," Diagnostics, vol. 13, no. 1, art. 72, 2023, doi: 10.3390/diagnostics13010072.

[11] "Advancing skin cancer detection through deep learning and fusion of patient metadata and skin lesion images," Scientific Reports, vol. 16, art. 1968, 2026, doi: 10.1038/s41598-025-26392-4.

[12] R. A. Krohling and A. G. C. Pacheco, "An attention-based mechanism to combine images and metadata in deep learning models applied to skin cancer classification," IEEE Journal of Biomedical and Health Informatics, vol. 25, no. 9, pp. 3554-3563, 2021, doi: 10.1109/JBHI.2021.3062002.

[13] W. Yin, J. Huang, J. Chen, and Y. Ji, "A study on skin tumor classification based on dense convolutional networks with fused metadata," Frontiers in Oncology, vol. 12, art. 989894, 2022, doi: 10.3389/fonc.2022.989894.

[14] J. Arevalo, T. Solorio, M. Montes-y-Gomez, and F. A. Gonzalez, "Gated multimodal units for information fusion," arXiv:1702.01992, 2017, doi: 10.48550/arXiv.1702.01992.

[15] P. Tschandl, C. Rosendahl, and H. Kittler, "The HAM10000 dataset, a large collection of multi-source dermatoscopic images of common pigmented skin lesions," Scientific Data, vol. 5, art. 180161, 2018, doi: 10.1038/sdata.2018.161.

[16] A. G. C. Pacheco, G. R. Lima, A. S. Salomao, et al., "PAD-UFES-20: A skin lesion dataset composed of patient data and clinical images collected from smartphones," Data in Brief, vol. 32, art. 106221, 2020, doi: 10.1016/j.dib.2020.106221.

[17] R. R. Selvaraju, M. Cogswell, A. Das, R. Vedantam, D. Parikh, and D. Batra, "Grad-CAM: Visual explanations from deep networks via gradient-based localization," International Journal of Computer Vision, vol. 128, pp. 336-359, 2019, doi: 10.1007/s11263-019-01228-7.

[18] N. V. Chawla, K. W. Bowyer, L. O. Hall, and W. P. Kegelmeyer, "SMOTE: Synthetic minority over-sampling technique," Journal of Artificial Intelligence Research, vol. 16, pp. 321-357, 2002, doi: 10.1613/jair.953.

[19] "Adaptive class rebalancing and spatial-aware metadata fusion for long-tailed skin lesion classification," Frontiers in Medicine, vol. 13, art. 1841617, 2026, doi: 10.3389/fmed.2026.1841617.



[20] Z. Hu, W. Mei, H. Chen, and W. Hou, "Multi-scale feature fusion and class weight loss for skin lesion classification," Computers in Biology and Medicine, vol. 176, art. 108594, 2024, doi: 10.1016/j.compbiomed.2024.108594.

[21] S. Ma, Y.-H. Chan, E. C. H. Ngai, and J. W. K. Ho, "Asynchronous and focal federated learning for skin lesion classification under local data scarcity and class imbalance," Computer Methods and Programs in Biomedicine, vol. 272, art. 109073, 2025, doi: 10.1016/j.cmpb.2025.109073.

[22] "Privacy preserving skin cancer diagnosis through federated deep learning and explainable AI," Scientific Reports, 2025, doi: 10.1038/s41598-025-19905-8.

[23] "A leakage-aware comparative benchmark of machine learning, deep learning, and transformer models for reliable leukemia detection," arXiv:2606.24944, 2026, doi: 10.48550/arXiv.2606.24944.

[24] S. Saha, M. M. Hemal, M. Z. A. Eidmum, and M. F. Mridha, "Knowledge distillation approach for skin cancer classification on lightweight deep learning model," Healthcare Technology Letters, 2025, doi: 10.1049/htl2.12120.

[25] P. Liu, W. Qian, H. Li, and J. Cao, "A relationship-aware mutual learning method for lightweight skin lesion classification," Digital Communications and Networks, vol. 11, no. 3, pp. 603-612, 2025, doi: 10.1016/j.dcan.2024.04.004.

[26] T. Tuncer, P. D. Barua, I. Tuncer, S. Dogan, and U. R. Acharya, "A lightweight deep convolutional neural network model for skin cancer image classification," Applied Soft Computing, vol. 162, art. 111794, 2024, doi: 10.1016/j.asoc.2024.111794.