# Melanoma Skin Cancer Detection by AI: Machine Learning to Explainable Multimodal Deep Learning techniques

Akbar Kushanoor*§, Sanjay K. Sahay†

*† *Department of CS & IS, Goa Campus, BITS Pilani, India*
§ *Staff Data Engineer, GE Aerospace, JFWTC- Bangalore,India*
{p20190079, ssahay}@goa.bits-pilani.ac.in

***Abstract*—Dermoscopy image analysis has emerged as a cornerstone problem in medical image computing and a rigorous benchmark for advancing artificial intelligence (AI) in healthcare. Despite significant progress from handcrafted feature-based methods to deep convolutional networks, encoder decoder segmentation frameworks, and ensemble learning, critical challenges related to generalization, interpretability, and clinical reliability still persist. Therefore, this study presents a comprehensive melanoma skin cancer detection using AI, from machine learning to explainable multimodal deep learning techniques for dermoscopic images. In contrast to existing reviews, we provide a critical meta-analysis of the design patterns underlying high-performing systems across benchmark datasets, including $PH^2$, HAM10000, PAD-UFES-20, Fitzpatrick-17k, and ISIC 2016-2019 challenges, and identify recurring architectural and training strategies that consistently drive performance improvements. Furthermore, we also discuss a structured taxonomy spanning supervised, self-supervised, convolutional neural networks, multimodal, and explainable AI paradigms, with particular emphasis on class imbalance handling, uncertainty estimation, and robustness to domain shifts. Finally, we outline emerging research directions toward trustworthy, generalizable, and clinically deployable AI systems for melanoma diagnosis.**

***Index Terms*—Melanoma, Skin Cancer, AI, Deep Learning, Explainable AI, Multimodal**

## I. Introduction

Dermoscopy is a widely used imaging modality for the assessment of pigmented skin lesions. However, it has recently posed significant challenges for AI because of substantial intra-class variability and subtle inter-class differences. In addition, dermoscopic images contain artifacts such as hair, rulers, gel marks, and bubbles, and datasets are typically highly imbalanced, with malignant lesions forming a minority of cases [5]. These factors complicate automated analysis and hinder the development of models that generalize across diverse clinical settings. Consequently, dermoscopy serves as a challenging yet valuable testbed for developing and benchmarking computer vision methods.

Driven by these challenges, the last two decades have seen rapid progress in the field of automated dermoscopic image analysis. Early approaches relied on handcrafted descriptors combined with classical classifiers, where features such as color histograms, texture measures, border irregularity indices, and asymmetry statistics were used with methods including k-nearest neighbors (kNN), support vector machines (SVMs), random forests, and shallow neural networks [1]. To capture more expressive local patterns, dictionary-based representations such as bag-of-features and fisher vectors were subsequently introduced [5]. Recently, a paradigm shift toward deep learning has transformed the field, with convolutional neural networks (CNNs) [59] and fully convolutional networks (FCNs) [47] becoming the backbone of most segmentation and classification systems [48], often integrated into ensemble frameworks with diverse architectures. Therefore, this paper focuses on the AI aspects of dermoscopic skin lesion analysis, emphasizing methodological design choices, performance trade-offs, and the recurring principles underlying successful systems. This paper discusses core machine learning and deep learning techniques, including image preprocessing and data partitioning, feature extraction, model architectures, segmentation and classification, class imbalance handling, and visual and textual explanation methods for model decisions. The objective is to provide readers with a unified and insight-driven synthesis of current AI-based approaches for dermoscopic skin lesion analysis.

The remainder of this paper is organized as follows. Section II reviews the preprocessing and data preparation techniques. Section III discusses feature extraction strategies. Section IV provides a comparative overview of classical and deep learning paradigms. Section V provides the training strategies and learning paradigms, and Section VI discusses the model explainability and analysis. Finally, Section VII concludes the paper with challenges for the detection of melanoma skin cancer using AI.

## II. Preprocessing and Data Preparation

Public dermoscopic datasets have been collected using different devices over extended periods and under varying acquisition conditions. Consequently, images often vary in terms of resolution, color characteristics, illumination, and the presence of artifacts. Preprocessing aims to reduce this variability and facilitate the learning of robust models that generalize across datasets and devices [6]. A prominent source of variation arises from differences in color and illumination. Color constancy and calibration methods attempt to transform device-dependent RGB values into stable representations. Some authors use physical or empirical calibration by imaging color charts or gray patches and fitting mappings into device-independent color spaces [11]. More commonly, in public datasets where such metadata are unavailable, algorithms such as Gray-World, Shades-of-Gray, or histogram-based equalization are applied to achieve color constancy. In this, Barata et al. showed that simple shades-of-gray normalization combined with contrast enhancement can improve cross-dataset performance for both handcrafted and deep features [9]. These observations are corroborated by subsequent surveys, which highlight the role of color normalization in improving robustness and cross-dataset generalization, particularly in heterogeneous dermoscopic datasets [5]. Artifact removal is an important preprocessing step because hair, frames, rulers, and bubbles can interfere with both segmentation and classification tasks. The classic DullRazor algorithm detects thin, dark structures using morphological filters, removes them via interpolation, and remains widely used as a baseline [13]. Building on this approach, subsequent methods have refined hair detection using multi-scale filtering, curve detection, and structure tensors, while applying more sophisticated inpainting techniques to fill occluded regions [14]. Although CNNs can tolerate some level of clutter, aggressive hair and frame removal can reduce false positives for lesion borders and stabilize the segmentation network training.

Beyond artifact removal, resizing and cropping strategies address the mismatch between arbitrary-sized dermoscopic images and fixed-sized inputs expected by most networks. A straightforward approach is to rescale the entire image to a common resolution with or without padding. In contrast, many classification pipelines focus on the lesion region by cropping around a ground truth or predicted segmentation mask. Studies have shown that training on tightly cropped lesions often improves classification performance compared to full images, particularly when applied consistently during both training and testing phases. To further enhance robustness, high-performing ISIC challenge submissions frequently employ multi-crop evaluation, where several crops at different scales or positions are classified independently, and their predictions are aggregated, thereby reducing the sensitivity to exact cropping and localization errors. Data partitioning and evaluation protocols also deserve careful attention. On small datasets such as $PH^2$, researchers commonly use k-fold cross-validation or repeated training and test splits [46]. The ISIC challenges provide predefined training, validation, and test sets, and strict adherence to these splits is crucial for a fair comparison and leader board validity. Robust evaluation protocols should ensure that data splits are patient-wise and that hyperparameters are tuned exclusively on validation data to prevent information leakage.

## III. Feature Extraction Strategies

Feature extraction is a key for the melanoma skin cancer detection using AI. In dermoscopy, the evolution of feature identification has led to broader developments in computer vision. In this, Barata et al. proposed a taxonomy that distinguishes between handcrafted features, dictionary-based representations, deep-learned features, and clinically inspired representations [5]. Handcrafted features describe lesions using explicitly defined parameters. Early systems employed color histograms in RGB, HSV, or Lab spaces, together with summary statistics such as channel-wise means and variances. Textures were captured using grey-level co-occurrence matrices, Gabor or wavelet filters, local binary patterns, or Laws masks. Shape and border descriptors, including compactness, eccentricity, fractal dimension, and radial distance measures, were used to capture irregular outlines, whereas simple asymmetry measures were used to compare intensity distributions across axes through the lesion centroid [16,17]. Such features can be computed directly from pixel data and often map onto intuitive visual properties, facilitating their interpretation and debugging. However, their expressiveness is inherently limited because they may fail to capture subtle cues and are sensitive to scale, rotation, and illumination, unless carefully normalized.

To address these limitations, dictionary-based representations increase flexibility. In a typical pipeline, each image is decomposed into a set of local patches sampled either on a regular grid or around salient points. These patches are described by low-level descriptors (e.g., Scale-Invariant Feature Transform (SIFT), color-SIFT, and gradient histograms), which are then clustered or used to learn a code book or sparse basis. The image is represented by statistics over codewords, such as bag-of-features histograms, vectors of locally aggregated descriptors, and Fisher vectors [21]. In dermoscopy, dictionary-based features have been investigated for both lesion-level classification and for detecting local structures, such as pigment networks or blue-white areas. Experiments have shown that they can outperform global handcrafted features, especially when combined with discriminative classifiers; however, their performance depends strongly on patch sampling, dictionary size, and pooling choices.

Recently, deep learning has fundamentally transformed feature representation learning by enabling end-to-end optimization. Deep-learned features dispense

with handcrafted designs and derive representations directly from data. In the simplest configuration, dermoscopic images are passed through a CNN pre-trained on ImageNet, and activations from intermediate layers are used as fixed feature vectors. Codella et al. combined such deep features with sparse coding and SVM classifiers and reported significant gains over purely handcrafted pipelines [20]. As larger annotated dermoscopy datasets have become available, fully end-to-end deep models have become feasible. Networks such as VGG, Inception, ResNet, DenseNet, and EfficientNet are routinely fine-tuned on dermoscopy images for classification, with the final layers adapted to the target classes. For segmentation, fully convolutional networks (FCNs) and U-Net variants learn dense feature maps that simultaneously encode semantics and localization [35]. Deep features are hierarchical and highly expressive; however, they are less interpretable, and their performance strongly depends on data availability and training strategies.

Complementary to these data-driven approaches, clinically inspired representations occupy orthogonal axes. Instead of representing images directly, they model intermediate visual concepts that dermatologists routinely use, such as pigment networks, streaks, dots and globules, regression structures, or the presence of specific colors. From an AI perspective, these can be treated as structured labels or auxiliary tasks. Kawahara et al. proposed FCNs that predict dermoscopic criteria at the pixel level and used the resulting maps as features for lesion-level classification [2]. Gonzalez-Diaz introduced a network that jointly predicts disease labels and dermoscopic attributes, effectively regularizing the representation space to respect expert knowledge [24]. Barata et al. reviewed a wide range of clinically inspired feature detectors and emphasized the potential of such representations to guide deep feature learning when combined with multitask objectives [5].

## IV. Taxonomy of AI Methods in Dermoscopic Skin Lesion Analysis

To provide a structured overview of the existing research, AI methods for dermoscopic skin lesion analysis can be broadly categorized into a taxonomy, as illustrated in Figure 1. These approaches fall into five major categories: feature-based methods, deep learning methods, hybrid and advanced models, explainable AI approaches, and multimodal learning frameworks. This taxonomy reflects the evolution of the field from traditional handcrafted pipelines to modern data-driven models, with an increasing emphasis on interpretability and the integration of complementary data sources.

### A. Traditional Machine Learning Models

Dermoscopic skin lesion analysis has traditionally relied on a two-stage pipeline, in which features are first extracted and subsequently fed into a classifier.

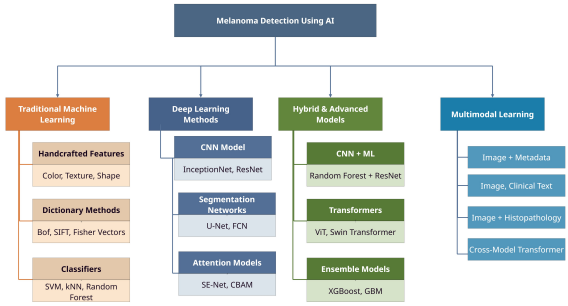


Fig. 1. Taxonomy of AI methods in dermoscopic skin lesion analysis.

Within this framework, this section provides a concise overview of the principal model families and highlights their continuing role as the baselines and components of hybrid systems. Among these methods, instance-based classifiers, such as k-nearest neighbors (kNN), are conceptually simple. Given a feature and distance metric, a new lesion is classified based on the labels of the most similar training data. However, kNN is primarily used as a baseline because its performance degrades in high-dimensional feature spaces and under severe class imbalances [3]. Despite these limitations, it remains useful for sanity checks and for exploring the structure of the feature space, for example, through neighbor retrieval and visualization. However, SVMs have been far more influential in this field because it construct hyperplanes that maximize the margin between classes and can be kernelized to represent the nonlinear decision boundaries. In dermoscopy, SVMs have been applied to handcrafted features, dictionary-based representations, and deep features extracted from pre-trained CNNs [8]. Accordingly, it performs well on moderate datasets with high-dimensional inputs and can be tuned to emphasize minority classes using class-weighted loss functions. Their main disadvantages are sensitivity to kernel and hyperparameter choices and computational cost for very large datasets. Tree-based methods, including decision trees, random forests, and gradient boosting, have also been widely adopted. Decision trees alone are prone to overfitting; however, random forests mitigate this by aggregating predictions from multiple trees trained on bootstrapped samples with random feature subsets. It is robust to mixed feature types and requires minimal feature scaling. Several studies have reported the strong performance of random forests on handcrafted feature sets, and their variable-importance measures provide some insight into which descriptors are the most discriminative [1]. Gradient boosting algorithms, such as XGBoost, extend this idea by sequentially adding trees to correct residual errors, often achieving higher accuracy at the cost of more careful regularization.

Shallow neural networks were among the earliest learning-based systems used for dermoscopy. In this, Binder et al. trained feed-forward networks on features derived from epiluminescence microscopy and demonstrated classification performance comparable to expert

dermatologists on modest-sized datasets [25]. Piccolo et al. evaluated a CAD system based on an artificial neural network and found that its sensitivity to melanoma was similar to that of a trained dermatologist, but with slightly lower specificity [26]. These studies showed that neural models could learn useful representations from dermoscopic features long before deep architectures became practical. Hybrid and ensemble classical models combine multiple feature sets and classifiers within a unified framework. Pathan et al. and others described systems in which different types of features are fed into separate SVMs or random forests, and the outputs are combined by majority voting or stacked into a meta-classifier [53]. Although such ensembles have largely been overtaken by deep learning ensembles in recent benchmarks, they remain useful as interpretable baselines and as components in systems where deep features are combined with handcrafted descriptors or metadata.

### B. Deep Learning Architectures

Deep learning has dramatically changed the way dermoscopic images are analyzed. Most recent studies use CNN architectures originally developed for large-scale natural image recognition and semantic segmentation, which are adapted to dermoscopy via transfer learning and task-specific modifications. For classification, many pipelines start with a CNN pre-trained on ImageNet. Early dermoscopy studies used AlexNet and VGG, which demonstrated the power of deep convolutional features but were relatively heavy and lacked architectural refinements such as residual connections. Inception architectures introduced modules that process feature maps with multiple kernel sizes in parallel, followed by $1 \times 1$ bottleneck convolutions to control the channel counts. Inception-v3 and Inception-v4 further refined this idea using factorized convolutions and deeper networks [28]. Residual networks (ResNets) added identity skip connections, allowing deeper networks to be trained and becoming a standard backbone in many domains, including dermoscopy.

DenseNets further utilize residual connections by concatenating feature maps from all previous layers into the input of each new layer, encouraging feature reuse and efficient gradient propagation. DenseNet-121 and DenseNet-161 are among the most commonly used backbones in ISIC submissions and often dominate the single-model baselines [39]. Xception and MobileNet replace standard convolutions with depthwise separable convolutions, reducing the parameter count and computational cost while maintaining performance [32]. These models are attractive for potential deployment on portable devices, although many challenge-winning methods still rely on heavier model architectures. Attention mechanisms have been integrated into CNNs to focus on informative channels and spatial locations. Squeeze-and-excitation (SE) blocks explicitly model inter-channel dependencies and reweight feature maps accordingly, improving accuracy with minimal parameter overhead [33]. Convolutional block attention modules (CBAM) combine channel and spatial attention to further refine feature maps [42]. Residual attention networks insert attention modules into residual architectures, enabling attention-aware features at multiple scales [41]. In dermoscopy, these modules have been used as drop-in improvements to standard backbones and as components of bespoke architectures for hierarchical diagnosis. For segmentation and pixel-wise tasks, FCNs and encoder–decoder architectures are the main choices. FCNs replace fully connected layers with $1 \times 1$ convolutions and use learnable upsampling layers to produce dense prediction maps [35]. U-Net introduced a symmetric encoder–decoder architecture with skip connections between corresponding resolution levels, allowing the network to combine high-level semantic information from the deep layers with fine spatial details from the shallow layers. Many U-Net variants have been proposed for skin lesion segmentation, including those with dilated convolutions, residual blocks, and multi-scale inputs.

DermoNet is an efficient segmentation architecture that is tailored for dermoscopy. It transforms DenseNet into a fully convolutional network with dense blocks in the encoder and a lightweight decoder connected via multi-scale skip connections. By reusing features across scales and restricting upsampling to feature maps from the most recent dense block, DermoNet achieves competitive or superior performance compared to FCN and U-Net baselines on ISBI 2016, ISBI 2017, and $PH^2$, while using fewer parameters and enabling faster inference [37]. Other architectures, such as DeepLab and RefineNet, incorporate additional techniques from semantic segmentation, including atrous convolutions and multi-scale context modules, and have been applied to dermoscopy segmentation. These architectures form the building blocks of most modern dermoscopy-analysis systems.

### C. Segmentation Networks

Segmentation networks are fundamental to medical image analysis, providing pixel-level predictions that enable the accurate delineation of pathological regions. Unlike image-level classification, these models generate dense prediction maps that precisely localize the lesion boundaries. In melanoma detection, segmentation facilitates robust region-of-interest extraction, reduces background artifacts, and improves both diagnostic accuracy and interpretability. Consequently, it remains a critical component of multimodal and explainable AI frameworks for reliable clinical decision support. Classical segmentation methods, including thresholding, region growing, clustering, and edge-based techniques, were widely used prior to the advent of deep learning [7]. Although computationally efficient, these approaches are

sensitive to noise, illumination variations, and fuzzy boundaries, often requiring problem-specific tuning. Segmentation continues to be a key component in many systems and is central to benchmarks such as ISIC 2016 and 2017.

In recent years, deep learning-based segmentation has largely replaced these classical techniques. U-Net and its derivatives have become the default choices. Vesal et al. and others extended U-Net by adding dilated convolutions at the bottleneck, multi-scale inputs, and deeper encoders, achieving Dice coefficients above 0.9 on ISBI 2017. Multistage FCN frameworks have also been proposed, in which an initial segmentation is refined by a second network or by combining outputs at multiple scales [35]. DermoNet represents an interesting point in the design space. By leveraging dense connectivity in the encoder and dense skip connections, it achieves accurate segmentation with a relatively small number of parameters. Quantitative evaluations on ISBI 2016, ISBI 2017, and $PH^2$ show that DermoNet matches or outperforms heavier architectures, such as FCN and U-Net, while being faster at inference. The choice of loss function is also important. Many studies use a combination of cross-entropy and overlap-based losses, such as Dice or Jaccard, to better handle the class imbalance between the lesion and background pixels. Some incorporate boundary-aware losses or regularizers to sharpen the edges. Post-processing steps, including conditional random fields (CRFs) and morphological operations, are sometimes applied to improve boundary smoothness and remove small spurious regions [38]. Segmentation maps are used in various downstream applications. In some pipelines, they serve solely to crop lesions before the classification. In other studies, shape descriptors or lesion areas were computed from the mask and used as additional features. Multi-task networks that jointly learn to segment and classify lesions have also been explored, with the intuition that segmentation provides a strong structural prior that benefits classification [57].

### *D. Classification and Ensembles*

Lesion classification is the primary end task for most dermoscopic CAD systems. Public benchmarks, including ISIC 2016–2019 and HAM10000, have spurred the rapid development of classification pipelines, culminating in ensembles that approach or exceed expert dermatologist performance under controlled conditions [58]. Single-model baselines typically fine-tune a CNN backbone on the target dataset. ResNet, DenseNet, Inception-v3, and EfficientNet are among the most commonly used neural network architectures. The models were initialized with ImageNet weights, and the last fully connected layer was replaced with one matching the number of lesion classes. Fine-tuning strategies vary; some authors freeze early layers and train only the last few blocks, whereas others fine-tune the entire network with a reduced learning rate. Menegola et al. systematically compared several transfer learning schemes and found that fine-tuning usually outperformed feature extraction-only approaches, particularly when combined with strong data augmentation.

Ensembles play a major role in the top-ranked entries of the ISIC challenges. In this, Gessert et al. used ensembles of up to 30 CNNs, including ResNet, DenseNet, SENet, and EfficientNet variants, trained at different input resolutions and with different augmentations [49]. The predictions were averaged or combined using simple voting. Harangi proposed smaller ensembles of 5–10 networks and demonstrated that ensembles consistently improve over individual models across different architectures [51]. These findings align with well-known results in machine learning: ensembles reduce variance and help mitigate overfitting, particularly when individual models are diverse. Input strategies differ across pipelines. Some methods operate on full images, possibly rescaled, whereas others use segmented or manually cropped lesions. Multiscale strategies, in which models observe both local details and broader contextual information, are particularly effective. For example, Valle et al. combined models trained on different input sizes and crop strategies and showed that this improves robustness across lesion types and acquisition settings [50]. Many pipelines also employ extensive test-time augmentation by averaging the predictions over multiple rotations, flips, and crops.

Recent models have incorporated metadata, such as patient age, sex, and lesion location, alongside image features. Akbar et al. concatenated CNN-derived features with embeddings of metadata in a multi-layer perceptron and reported modest but consistent gains in AUC and balanced accuracy on ISIC 2019 [61]. Incorporating clinical metadata raises questions regarding how to weight image and non-image signals and how to handle missing or noisy metadata. A recurring challenge is severe class imbalance. In multi-class settings, such as ISIC 2018 and 2019, common benign lesions vastly outnumber malignant or rare lesion types. Many researchers have addressed this issue using class-weighted loss functions, focal loss, or oversampling of minority classes. Gessert et al. used loss weighting proportional to the inverse class frequency combined with balanced mini-batches and augmentation, leading to a higher sensitivity for underrepresented classes [48]. Others have experimented with SMOTE- or GAN-based augmentations, although the impact of generative augmentation on the final performance remains debatable.

Comparisons with human experts provide useful reality checks. Akbar et al. trained an Inception-v3 network on a large proprietary dataset and showed that its melanoma classification ROC curve overlapped with that of a group of dermatologists on a test set of clinical and dermoscopic images [59]. Brinker et al. evaluated a CNN trained on dermoscopy images against 157 dermatologists and found that the network achieved similar or higher sensitivity for melanoma at comparable specificity levels [55]. These results depend strongly on the

case mix and evaluation design but illustrate that deep models can be competitive with experts under certain conditions.

## V. Training Strategies and Learning Paradigms

Beyond architecture and data, training strategy plays a central role in determining model performance. Therefore, this section discusses several important aspects: data augmentation, class-imbalance handling, transfer learning, and semi/weak supervision. Data augmentation is almost universally used to reduce overfitting and improve generalization. Common transformations include rotation, flip, scaling, translation, and color jittering. Vasconcelos et al. systematically evaluated various augmentations and found that aggressive geometric augmentation can significantly improve the performance on small dermoscopy datasets [5]. Color augmentation, such as random brightness, contrast, and hue shifts, helps make models less sensitive to illumination changes. Some studies have applied techniques such as mixup or cutout, although their use in dermoscopy remains more limited than in natural-image tasks. Class imbalance is a persistent issue. When malignant lesions represent a small fraction of the training data, naive cross-entropy training can yield models that largely ignore minority classes. Several strategies have been explored. Loss weighting assigns higher penalties to errors in underrepresented classes, typically with weights that are inversely proportional to class frequencies. Focal loss down-weights easy examples and focuses on difficult cases and has been adopted in recent dermoscopy models for heavily skewed settings. Data-level strategies include oversampling minority classes, either by duplicating examples or using SMOTE-like interpolation, and undersampling majority classes. These approaches can reduce bias but may introduce overfitting to duplicated samples. Many high-performing systems combine several techniques.

Transfer learning is almost always used. Starting from ImageNet-pretrained weights speeds up convergence and improves performance compared to training from scratch, particularly when training data are limited. Menegola et al. compared feature extraction-only schemes (freezing convolutional layers and training only the classifier) with partial and full fine-tuning and concluded that fine-tuning at least the deeper convolutional blocks consistently improves the classification performance on dermoscopy datasets [52]. Adegun and Viriri reviewed a wide range of studies and confirmed that transfer learning is a de facto standard in the field [4]. Semi-supervised and weakly supervised learning have begun attracting attention. One line of work uses unlabeled dermoscopy images to pre-train models via self-supervised tasks, such as predicting image rotations or solving contrastive learning objectives, before fine-tuning on small labeled datasets. Another approach leverages weak labels, such as image-level tags for dermoscopic structures or clinical descriptors, to train models that predict both the diagnosis and attributes. Multitask learning is closely related. Networks that jointly predict the lesion class and dermoscopic criteria can benefit from shared representations, with attribute prediction serving as a form of regularization [24]. Some studies have also combined segmentation and classification in a single model by sharing an encoder between the segmentation decoder and classification head. Empirical evidence suggests that multitask training can improve data efficiency and stabilize training, although the optimal choice of tasks and loss weighting remains an open question.

## VI. Explainability and Model Analysis

Deep models achieve strong performance but are often criticized for their opacity. In dermoscopy, where visual patterns are rich and domain knowledge is well developed, there is particular interest in understanding what models have learned and whether their decisions are based on plausible image cues. Class Activation Mapping (CAM) and gradient-based variants (Grad-CAM, Grad-CAM++) have been widely used to visualize the regions of an image that contribute most to a network's prediction. When applied to dermoscopy, these methods typically highlight the lesion area and sometimes emphasize specific structures such as pigment networks or darker regions [62]. While CAMs offer a coarse localization signal, their interpretability depends on the alignment between the highlighted regions and known dermoscopic structures.

Attention mechanisms provide another explanation. In residual attention networks and CBAM-augmented CNNs, attention maps can be visualized to show where the model focuses spatially or which feature channels are emphasized [33]. Barata et al. proposed a deep attention model for the hierarchical diagnosis of skin lesions by inserting attention modules at different levels of a DenseNet-based architecture [40]. They reported that attention maps often concentrate on regions consistent with human expectations and that the hierarchical formulation improves balanced accuracy on ISIC 2017. Clinically inspired auxiliary tasks also serve explanatory purposes. If a model explicitly predicts dermoscopic criteria such as pigment networks, streaks, or regression areas, internal representations can be probed in terms of these higher-level concepts [60]. Concept-based explanation methods, such as Testing with Concept Activation Vectors (TCAV), could, in principle, quantify the contribution of each attribute to decisions, although this line of work remains relatively unexplored in dermoscopy.

Content-based image retrieval (CBIR) systems built using deep embeddings offer an alternative explanation. Instead of highlighting pixels, they retrieve similar lesions from a reference database along with their diagnoses. From a machine learning standpoint, these systems rely on metric learning or embedding losses to

ensure that visually or diagnostically similar lesions are close in feature space. Overall, explainability methods in dermoscopy remain in an exploratory phase, and most studies provide qualitative examples rather than systematic user studies or quantitative measurements.

## VII. Conclusion and Challenges

In this paper, we present a comprehensive review of artificial intelligence techniques for dermoscopic skin lesion analysis, covering the full pipeline from preprocessing and feature representation to segmentation, classification, and explainability. Although deep learning approaches, particularly convolutional and attention-based architectures, have achieved remarkable performance, several critical challenges continue to limit their clinical applicability. A primary concern is the domain shift issue, wherein models trained on specific datasets exhibit degraded performance when applied to images from different acquisition settings, devices, or patient populations. Addressing this limitation requires a systematic investigation of domain generalization and adaptation techniques. Additionally, model robustness and calibration remain insufficiently explored, as many existing systems produce overconfident predictions when confronted with out-of-distribution inputs or imaging artifacts. Therefore, incorporating uncertainty estimation and robust optimization strategies is essential for reliable deployment. Another fundamental limitation is the scarcity of data and the high cost of annotation. Although publicly available datasets have enabled substantial progress, they remain relatively small and imbalanced. Future studies should increasingly leverage self, semi, and weakly supervised learning to exploit large volumes of unlabeled data. Multimodal learning represents a promising, yet comparatively underexplored, direction. Integrating dermoscopic images with complementary sources, such as clinical metadata, patient history, textual reports, and histopathological information, has the potential to significantly enhance diagnostic performance and model reliability. However, effective fusion strategies and standardized evaluation protocols for multimodal systems remain open research challenges. Finally, explainability and user-centered evaluations are crucial for clinical translation. Beyond conventional performance metrics, it is necessary to develop interpretable models whose decision-making processes align with dermatological reasoning and rigorously assess their impact on clinician trust, calibration, and diagnostic outcomes. In summary, although dermoscopic skin lesion analysis has emerged as a compelling benchmark for AI in medical imaging, future progress will depend on advancing generalization, robustness, data efficiency, multimodal integration, and trustworthy explainability, supported by closer collaboration between machine learning researchers and clinical experts.

## References


[1] K. Korotkov and R. Garcia, "Computerized analysis of pigmented skin lesions: a review," *Artificial Intelligence in Medicine*, vol. 56, no. 2, pp. 69–90, 2012.

[2] J. Kawahara and G. Hamarneh, "Seven-Point Checklist and Skin Lesion Classification Using Multitask Fully Convolutional Networks," *IEEE Journal of Biomedical and Health Informatics*, vol. 23, no. 2, pp. 538–546, 2019.

[3] R. B. Oliveira, J. P. Papa, A. S. Pereira, and J. M. R. S. Tavares, "Computational methods for pigmented skin lesion classification in images: review and future trends," *Neural Computing and Applications*, vol. 29, no. 3, pp. 613–636, 2018.

[4] A. Adegun and S. Viriri, "Deep learning techniques for skin lesion analysis and melanoma cancer detection: a survey of state-of-the-art," *Artificial Intelligence Review*, vol. 53, pp. 1–73, 2020.

[5] C. Barata, M. E. Celebi, and J. S. Marques, "A Survey of Feature Extraction in Dermoscopy Image Analysis of Skin Cancer," *IEEE Journal of Biomedical and Health Informatics*, vol. 23, no. 3, pp. 1109–1120, 2019.

[6] M. E. Celebi, T. Mendonça, and J. S. Marques, *Dermoscopy Image Analysis*. CRC Press, 2015.

[7] M. E. Celebi, Q. Wen, S. Hwang, H. Iyatomi, X. Chen, and G. Schaefer, "A state-of-the-art survey on lesion border detection in dermoscopy images," *Dermoscopy Image Analysis*, pp. 97–129, 2015.

[8] N. K. Mishra and M. E. Celebi, "An overview of melanoma detection in dermoscopy images using image processing and machine learning," *arXiv preprint arXiv:1601.07843*, 2016.

[9] C. Barata, M. E. Celebi, and J. S. Marques, "Improving dermoscopy image classification using color constancy," *IEEE Journal of Biomedical and Health Informatics*, vol. 19, no. 3, pp. 1146–1152, 2015.

[10] Y. Van Haeghen, J. M. A. D. Naeyaert, and I. Lemahieu, "An imaging system with calibrated color image acquisition for use in dermatology," *IEEE Transactions on Medical Imaging*, vol. 19, no. 7, pp. 722–730, 2000.

[11] C. Grana, G. Pellacani, and S. Seidenari, "Practical color calibration for dermoscopy applied to a digital epiluminescence microscope," *Skin Research and Technology*, vol. 11, no. 4, pp. 242–247, 2005.

[12] J. Quintana, R. Garcia, and L. Neumann, "A novel method for color correction in epiluminescence microscopy," *Computerized Medical Imaging and Graphics*, vol. 35, no. 8, pp. 646–652, 2011.

[13] T. Lee, V. Ng, R. Gallagher, A. Coldman, and D. McLean, "Dullrazor: A software approach to hair removal from images," *Computer Methods and Programs in Biomedicine*, vol. 27, no. 6, pp. 533–543, 1997.

[14] Q. Abbas, M. E. Celebi, and I. F. García, "Unsupervised skin lesion border detection via two-dimensional image analysis," *Computer Methods and Programs in Biomedicine*, vol. 104, no. 3, pp. e1–e15, 2011.

[15] Q. Abbas, M. E. Celebi, and I. G. Fondon, "Hair removal methods: a comparative study for dermoscopy images," *Biomedical Signal Processing and Control*, vol. 6, no. 4, pp. 395–404, 2013.

[16] E. Claridge and P. N. Hall, "Shape analysis for classification of malignant melanoma," *Journal of Biomedical Engineering*, vol. 14, no. 3, pp. 229–234, 1992.

[17] V. T. Y. Ng, B. Y. M. Fung, and T. K. Lee, "Determining the asymmetry of skin lesion with color and shape information," in *Proc. EMBS*, 1999.

[18] P. N. Hall, E. Claridge, and J. D. Morris Smith, "Computer screening for early detection of melanoma: is there a future?," *British Journal of Dermatology*, vol. 132, no. 3, pp. 325–338, 1995.

[19] C. Barata, J. S. Marques, and J. Rozeira, "A system for the detection of pigment network in dermoscopy images using directional filters," *IEEE Transactions on Biomedical Engineering*, vol. 59, no. 10, pp. 2744–2754, 2012.

[20] N. C. F. Codella, J. Cai, M. Abedini et al., "Deep learning, sparse coding, and SVM for melanoma recognition in dermoscopy images," in *MICCAI Workshop on Machine Learning in Medical Imaging*, pp. 118–126, 2015.

[21] L. Yu, H. Chen, Q. Dou, J. Qin, and P. A. Heng, "Fisher vector encoding of deep features for dermoscopy image classification," *Computer Vision and Image Understanding*, vol. 169, pp. 118–129, 2018.

[22] G. Albuquerque et al., "Using autoencoders to improve classification of skin lesions," *Computer Methods and Programs in Biomedicine*, 2017.

[23] J. Kawahara, S. Daneshvar, G. Argenziano, and G. Hamarneh, "Fully convolutional neural networks to detect clinical dermoscopic features," *IEEE Journal of Biomedical and Health Informatics*, vol. 23, no. 2, pp. 578–585, 2019.

[24] I. Gonzalez-Diaz, "DermaKNet: Incorporating the knowledge of dermatologists to convolutional neural networks for skin lesion diagnosis," *IEEE Journal of Biomedical and Health Informatics*, vol. 23, no. 2, pp. 547–559, 2019.

[25] M. Binder, A. Steiner, M. Schwarz et al., "Application of an artificial neural network in epiluminescence microscopy pattern analysis of pigmented skin lesions: a pilot study," *British Journal of Dermatology*, vol. 130, no. 4, pp. 460–465, 1994.

[26] D. Piccolo, A. Ferrari, K. Peris, R. Daidone, B. Ruggeri, and S. Chimenti, "Dermoscopic diagnosis by a trained clinician vs. a clinician with minimal dermoscopy training vs. computer-aided diagnosis," *British Journal of Dermatology*, vol. 147, no. 3, pp. 481–486, 2002.

[27] A. Krizhevsky, I. Sutskever, and G. E. Hinton, "ImageNet classification with deep convolutional neural networks," in *Proc. NIPS*, pp. 1097–1105, 2012.

[28] C. Szegedy, W. Liu, Y. Jia et al., "Going deeper with convolutions," in *Proc. CVPR*, pp. 1–9, 2015.

[29] C. Szegedy, S. Ioffe, V. Vanhoucke, and A. A. Alemi, "Inception-v4, Inception-ResNet and the impact of residual connections on learning," in *Proc. AAAI*, 2017.

[30] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *Proc. CVPR*, pp. 770–778, 2016.

[31] G. Huang, Z. Liu, L. Van Der Maaten, and K. Q. Weinberger, "Densely connected convolutional networks," in *Proc. CVPR*, pp. 4700–4708, 2017.

[32] F. Chollet, "Xception: Deep learning with depthwise separable convolutions," in *Proc. CVPR*, 2017.

[33] J. Hu, L. Shen, and G. Sun, "Squeeze-and-excitation networks," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 42, no. 8, pp. 2011–2023, 2020.

[34] M. Tan and Q. V. Le, "EfficientNet: Rethinking model scaling for convolutional neural networks," in *Proc. ICML*, 2019.

[35] J. Long, E. Shelhamer, and T. Darrell, "Fully convolutional networks for semantic segmentation," in *Proc. CVPR*, pp. 3431–3440, 2015.

[36] O. Ronneberger, P. Fischer, and T. Brox, "U-Net: Convolutional networks for biomedical image segmentation," in *Proc. MICCAI*, pp. 234–241, 2015.

[37] S. Baghersalimi, B. Bozorgtabar, P. Schmid-Saugeon, H. K. Ekenel, and J. P. Thiran, "DermoNet: densely linked convolutional neural network for efficient skin lesion segmentation," *EURASIP Journal on Image and Video Processing*, vol. 2019, no. 71, 2019.

[38] L. C. Chen, Y. Zhu, G. Papandreou, F. Schroff, and H. Adam, "Encoder-decoder with atrous separable convolution for semantic image segmentation," in *Proc. ECCV*, 2018.

[39] C. Barata and J. S. Marques, "Deep Learning for Skin Cancer Diagnosis With Hierarchical Architectures," in *Proc. ISBI*, pp. 841–845, 2019.

[40] C. Barata, J. S. Marques, and M. E. Celebi, "Deep Attention Model for the Hierarchical Diagnosis of Skin Lesions," in *Proc. CVPR Workshops*, 2019.

[41] F. Wang, M. Jiang, C. Qian et al., "Residual attention network for image classification," in *Proc. CVPR*, pp. 3156–3164, 2017.

[42] S. Woo, J. Park, J. Y. Lee, and I. S. Kweon, "CBAM: Convolutional block attention module," in *Proc. ECCV*, pp. 3–19, 2018.

[43] D. Gutman, N. C. F. Codella, M. E. Celebi et al., "Skin lesion analysis toward melanoma detection: A challenge at the 2016 ISBI," *arXiv preprint arXiv:1605.01397*, 2016.

[44] N. C. F. Codella, D. Gutman, M. E. Celebi et al., "Skin lesion analysis toward melanoma detection: A challenge at the 2017 International Symposium on Biomedical Imaging," in *Proc. ISBI*, pp. 168–172, 2018.

[45] "ISIC 2019: Skin Lesion Analysis Towards Melanoma Detection," https://challenge2019.isic-archive.com/, 2019.

[46] T. Mendonça, P. M. Ferreira, J. S. Marques et al., "PH2—A dermoscopic image database for research and benchmarking," in *Proc. EMBC*, pp. 5437–5440, 2013.

[47] 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, pp. 180161, 2018.

[48] N. Gessert, M. Nielsen et al., "Skin lesion diagnosis using ensembles, unscaled multi-crop evaluation and loss weighting," in *MICCAI Workshop on ISIC 2018*, 2018.

[49] N. Gessert, T. Sentker, F. Madesta et al., "Skin lesion classification using ensembles of multi-resolution EfficientNets with metadata," in *MICCAI Workshop on ISIC 2019*, 2019.

[50] E. Valle, M. Fornaciali, A. Menegola, and S. Avila, "Data, depth, and design: Learning reliable models for skin lesion analysis," *Neurocomputing*, 2019.

[51] B. Harangi, "Skin lesion classification with ensembles of deep convolutional neural networks," *Journal of Biomedical Informatics*, vol. 86, pp. 25–32, 2018.

[52] A. Menegola, M. Fornaciali, R. Pires, F. V. Bittencourt, S. Avila, and E. Valle, "Knowledge Transfer for Melanoma Screening with Deep Learning," in *Proceedings of the IEEE International Symposium on Biomedical Imaging (ISBI)*, pp. 297–300, 2017.

[53] S. Pathan, K. G. Prabhu, and P. C. Siddalingaswamy, "Techniques and algorithms for computer aided diagnosis of pigmented skin lesions: A review," *Biomedical Signal Processing and Control*, vol. 39, pp. 237–262, 2018.

[54] A. Esteva, B. Kuprel, R. A. Novoa et al., "Dermatologist-level classification of skin cancer with deep neural networks," *Nature*, vol. 542, no. 7639, pp. 115–118, 2017.

[55] T. J. Brinker, A. Hekler, A. H. Enk et al., "Skin cancer classification performance of a convolutional neural network vs. dermatologists," *European Journal of Cancer*, vol. 119, pp. 11–17, 2019.

[56] N. C. F. Codella, V. Rotemberg, P. Tschandl et al., "Collaborative human-AI melanoma diagnosis using content-based image retrieval," in *MICCAI Workshop on Understanding and Interpreting Machine Learning in Medical Image Computing*, 2018.

[57] Z. Mirikharaji et al., "A survey on deep learning for skin lesion segmentation," *Medical Image Analysis*, 2023.

[58] M. A. Kassem et al., "Machine Learning and Deep Learning Methods for Skin Cancer Classification and Detection: A Review," *Medical Imaging and Health Informatics*, 2021.

[59] A. Kushanoor and S. K. Sahay, "An Investigation of Deep Learning Techniques for the Robust Detection of Melanoma Cancer," in *Emerging Technology and Sustainable Solutions*, pp. 138–154, 2026.

[60] A. Kushanoor and S. K. Sahay, "MSF-CAM: Meta-Aware Multi-Scale Focused Class Activation Mapping for Trustworthy Skin Lesion Diagnosis," in *2025 IEEE 37th International Conference on Tools with Artificial Intelligence (ICTAI)*, pp. 678-685, 2025.

[61] A. Kushanoor and S. K. Sahay, "AI-Driven Dermatology: Enhancing Melanoma Detection with XAI and Deep Learning," in *Computer Vision, Pattern Recognition, Image Processing, and Graphics*, pp. 206–216, Springer Nature Switzerland, 2026.

[62] A. Kushanoor and S. K. Sahay, "Explainable Skin Cancer Detection via Hybrid CNN and Adaptive Post-hoc Explanations," in *Computer Vision, Pattern Recognition, Image Processing, and Graphics*, pp. 217–228, Springer Nature Switzerland, Cham, 2026.