Published by :
https://www.ijert.org/
An International Peer-Reviewed Journal

International Journal of Engineering Research & Technology (IJERT)
ISSN: 2278-0181
Vol. 15 Issue 10 , October - 2026

Skin Lesion Classification Using Machine Learning and Deep Learning: A Narrative Review of Datasets, Methods, Clinical Evidence, and Open Challenges

Prashansa D. Choksi | Assistant Professor,
Gyanmanjari Innovative University, Bhavnagar, Gujarat India

Abstract - Skin cancer is among the most prevalent cancers in the world, and early detection of melanoma is the best strategy to lower the associated mortality. Computer-aided diagnosis of skin lesions evolved from hand-crafted feature pipelines based on the ABCD rule and texture descriptors into deep neural networks achieving dermatologists' performance on benchmark-controlled tasks. In this review, we provide a full overview of the automated skin lesion classification pipeline: from the clinical background, over public skin lesion image datasets and challenges, and the early approach of hand-crafted computer vision descriptors over modern state-of-the-art deep learning architectures, vision transformers, self-supervised, and foundation models, multimodal skin lesion analysis explainability uncertainty estimation, down to reader studies or other comparative analyses between algorithms and clinicians. We provide a critical reflection on how far benchmark performance is from translational readiness, and address problems of dataset bias and under-representation of skin types; spurious correlations and artefacts; class imbalance; domain shift; reporting quality; privacy; and regulation. Finally, we conclude with a shared vision for future research in the field of skin cancer detection including the curation of human-labelled representative datasets; use of prospective validation studies; human-AI partnership systems; private training data; and responsible, explainable models. This is a narrative review article and not a systematic review and meta-analysis. We have described results of articles at the study level, and note that differences across performance metrics should be put into context of differences across datasets, reference tasks, and validation approaches.

Keywords: skin lesion classification; melanoma; dermoscopy; deep learning; convolutional neural networks; vision transformers; explainable AI; dataset bias; computer-aided diagnosis

1. INTRODUCTION

Skin cancer is the most common cancer in fair-skinned groups. Melanoma makes up only a small percentage of skin cancers but accounts for most skin-cancer-related deaths. In 2020, over 320,000 new cases of melanoma were projected to occur worldwide, leading to nearly 57,000 deaths [1]; in the United States in 2024 100 000 new cases of invasive melanoma are expected [2]. The prognosis correlates Really with tumour thickness at diagnosis: localised melanoma has a five-year survival rate over 99%, but this drops Greatly once the disease has spread [2]. These statistics make early identification a clinical priority, but this is challenging: benign moles, seborrhoeic keratoses, and other growths can look similar to melanoma, and even experts often disagree [3]. Machine Learning has played a prompt role in healthcare like cervical cancer prediction, also in social engineering detection. [75],[76].

Dermoscopy, a magnifying instrument with either polarised or immersion illumination, which eliminates the reflection of the skin surface and so boosts the visualisation of subsurface structures, increases the diagnostic performance of the naked eye when used by trained dermatologists [3]. Well-designed algorithms like the ABCD rule [4] or the pattern-analysis and checklist algorithms used at the Internet consensus meeting [5] help the expert to reason with the data; these algorithms can be learned through training, but are inherently subjective. This explained the interest in automated analysis for decades. Early CAD systems made use of hand-crafted features such as asymmetry, border irregularity, colour variegation, or texture, fed into conventional classifiers like support vector machines, decision trees, or k nearest neighbours [6] [8].

The deep learning revolution put an end to all this. Representation learning with multilayer networks [9], accelerated by superbly labelled big data [10], [11] and GPU acceleration, made end-to-end learning from pixels possible. In 2017, Esteva et al. published that a single CNN trained on 129,450 clinical images gained dermatologist performance for keratinocyte carcinoma and melanoma classification [12]. Haenssle et al. published that a CNN outperformed a majority of 58 dermatologists in dermoscopic

(This work is licensed under a Creative Commons Attribution 4.0 International License.)

International Journal of Engineering Research & Technology (IJERT)
ISSN: 2278-0181
Vol. 15 Issue 10 , October - 2026

melanoma recognition [13]. These studies and others sparked a barrage of publications, public challenges and commercial interests.

There are already a number of publications in parts of this domain. Takiddin et al. provided a scoping review of AI for skin cancer detection [14], Wu et al. reviewed deep learning for skin cancer classification [15], and Dick et al. performed a meta-analysis of CAD accuracy for melanoma [16]. Our review complements these in three aspects. First, it encompasses the entire pipeline from data to deployment and is not limited to model architectures; secondly, it offers a synthesis of the fast-growing body of transformers, self-supervised learning, and foundation models; finally, it highlights issues of fairness, robustness, and clinical validation impacting actual patient benefit.

Scope and organisation. This is a narrative review that spans from the earliest published CAD systems to recent deep learning publications. Section 2 presents the clinical background and imaging ways. Section 3 introduces the public datasets. Section 4 presents considerations on preprocessing segmentation augmentation, and class imbalance. Section 5 reviews CNN-based classifiers, Section 6 covers transformers and self-supervised and foundation models, and Section 7 addresses multimodal fusion. Section 8 considers explainability and uncertainty. Section 9 synthesises comparisons with clinicians. Sections 10 and 11 discuss challenges and future directions, and Section 12 concludes.

1.1. Review approach and evidence selection

This article is a narrative review of machine-learning and deep-learning approaches to skin lesion classification. It synthesises a range of selected studies published in peer-reviewed literature, descriptions of public datasets, challenge reports, clinical-reader validation studies and reporting or evaluation standards quoted in the article. The review includes information on classical computer-aided diagnosis approaches, convolutional neural networks, vision transformers, self-supervised learning, foundation models, multimodal approaches explainability fairness, robustness, and clinical implementation. The synthesis is qualitative; no pooled effect estimate or new meta-analysis is provided.

Table 1A. Positioning of this review relative to cited reviews.

Review / year	Review type and focus	Coverage indicated by cited source/title	Implication for the present review
Takiddin et al. [14] (2021)	Scoping review of AI for skin cancer detection	Provides broad mapping of AI applications and evidence landscape.	This manuscript should differentiate itself through an integrated pipeline-to-deployment synthesis and explicit attention to fairness, robustness, foundation models, and clinical readiness.
Wu et al. [15] (2022)	Systematic review of deep learning for skin cancer classification	Directly overlaps with deep-learning architectures and classification studies.	The present review must make its added value explicit by connecting datasets, validation design, human–AI evidence, and deployment barriers rather than repeating architecture summaries.
Dick et al. [16] (2019)	Meta-analysis of computer-aided melanoma diagnostic accuracy	Provides quantitative synthesis of diagnostic-accuracy evidence available to that review.	This manuscript is not a meta-analysis; it should avoid pooled-accuracy claims and explain why cross-study estimates are not directly comparable.
Present manuscript	Narrative review across data, methods, clinical evidence, and translation	Combines dataset limitations, CNNs, transformers/self-supervision/foundation models, multimodal fusion, explainability, fairness, and deployment considerations.	Its contribution is integrative and qualitative; a reproducible search log and more structured study-level evidence tables would strengthen credibility.

2. CLINICAL BACKGROUND AND IMAGING MODALITIES

2.1 Lesion categories

Skin lesion classification is usually posed as either a binary task (malignant versus benign, or melanoma versus non-melanoma) or a multiclass task. The seven diagnostic categories of the HAM10000 dataset are commonly used: melanocytic nevus melanoma benign keratosis-like lesions, basal cell carcinoma, actinic keratosis and intraepithelial carcinoma, dermatofibroma, and vascular lesions [22]. Larger collections add squamous cell carcinoma and an "unknown" class [23], [24]. Clinical practice involves

(This work is licensed under a Creative Commons Attribution 4.0 International License.)

International Journal of Engineering Research & Technology (IJERT)
ISSN: 2278-0181
Vol. 15 Issue 10 , October - 2026

hundreds of diagnoses, including inflammatory and infectious conditions, and dermatology-wide classifiers have been trained on tens of conditions or more [12], [58]. The difficulty varies: distinguishing nevi from melanoma is clinically subtle, while distinguishing vascular lesions is comparatively easy, which is one reason that aggregate accuracy can hide poor performance on clinically important minority classes.

2.2 Imaging modalities

There are three main image methods in the literature. Macro (clinical) photographs taken with a camera or smartphone are cheap and easy to obtain but have a high variability of lighting, distance, and background. There have been evaluations of smartphone-based applications in a systematic review, and a review of diagnostic-accuracy studies identified limited and biased evidence for their accuracy in risk layering [18]. Dermoscopic images give standardised, magnified views of subsurface pigment networks and structures in the skin, and most benchmark datasets are in this modality [3], [22]. Whole-slide histopathology images are the diagnostic gold standard, and deep learning techniques have also been used here, where one study showed improved performance over eleven pathologists in classifying melanoma and nevi slides [17]. Other ways are available including reflectance confocal microscopy, total-body photography and optical coherence tomography, but these are less represented in public datasets.

2.3 Why the task is hard

The images of skin lesions are very variable within class and similar across class, with a number of other nuisance factors like hairs rulers ink marks, air bubbles, vignetting, and colour calibration differences. Positivity (prevalence) is extremely skewed: in real-life data benign nevi are a lot more prevalent than melanoma, while curated datasets are biased to oddities and biopsy-derived diagnoses. Label quality (accuracy) is also variable: histopathology is the gold standard, but most images in public datasets are granted by expert consensus or subsequent diagnoses rather than biopsy [22].

3. DATASETS AND BENCHMARKS

Public datasets and challenges have been central to progress. Table 1 summarises the principal resources.

Table 2. Principal public skin lesion datasets.

Dataset	Year	Modality	Size (approx.)	Notes
PH2	2013	Dermoscopy	200 images	Small; segmentation and clinical annotations [19]
ISIC 2016	2016	Dermoscopy	900 train / 379 test	Binary melanoma classification; segmentation [20]
ISIC 2017	2017	Dermoscopy	2,000 train / 150 val / 600 test	Melanoma, seborrhoeic keratosis, nevus; segmentation [21]
HAM10000 / ISIC 2018	2018	Dermoscopy	10,015 images	Seven classes; lesion-level duplicates exist [22], [23]
BCN20000	2019	Dermoscopy	19,424 images	Includes difficult locations such as nails and mucosa [24]
ISIC 2019	2019	Dermoscopy	25,331 images	Eight classes plus out-of-distribution class; built from HAM10000, BCN20000, MSK [23], [24]
SIIM-ISIC 2020	2020	Dermoscopy + metadata	33,126 images (2,056 patients)	Patient-centric; strong class imbalance [25]
PAD-UFES-20	2020	Smartphone clinical + metadata	2,298 lesion samples	Six classes; patient clinical data [26]
Fitzpatrick17k	2021	Clinical	16,577 images	114 conditions with skin-type labels [27]

(This work is licensed under a Creative Commons Attribution 4.0 International License.)

International Journal of Engineering Research & Technology (IJERT)
ISSN: 2278-0181
Vol. 15 Issue 10 , October - 2026

Dataset	Year	Modality	Size (approx.)	Notes
Diverse Dermatology Images (DDI)	2022	Clinical	656 images	Biopsy-proven; balanced skin tones [28]

3.1 The ISIC archive and challenges

The ISIC (International Skin Imaging Collaboration) has run annually challenges since 2016, with increasing data volumes and multi-task developments of lesion segmentation, dermoscopic attribute detection and disease diagnosis [20], [21], [23]. In 2016, a challenge comparing algorithms to dermatologists showed the best algorithms were competitive in a standardised limited melanoma detection task [66]. HAM10000 was a dataset of 10,015 dermoscopy images of 7 diagnostic categories, which has become the standard benchmark [22]; subsequent challenges integrated more diverse multi-centre data [24], and the 2020 challenge introduced patient-level information to mimic clinical decision making [25].

3.2 Limitations of widely used datasets

Limitations but, these datasets are not without their drawbacks. For example, they first and foremost feature images of patients with fair skin in Europe, North America, and Australiadark skin (Fitzpatrick V and VI) is severely under-represented in dermatology datasets and skin-type data is often absent [29], [30]. Other limitations include in reality multiple images of the same lesion exists in many datasets, meaning that random splits at the image-level often means the information being "leaked" between training and test subsets, this way artificially improving accuracy [22]. Collection artefacts can also shortcut detection, for instance 'surgical skin markings' around lesions on skin lesion photographs increase the chance of a false-positive melanoma prediction by a CNN [31]. Clinical image datasets assembled from web search or image atlases may not represent the case mix or image quality found in practice [27], [28]. Developmental resources like Fitzpatrick17k [27] and DDI [28] help to address skin-tone bias, but are small in size compared to dermoscopic benchmarks, and the skin-type labels in Fitzpatrick17k are derived from crowd annotation rather than clinical assessment. Headlines on a single benchmark are because of this poor measures of real-world performance. Careful studies must include per-patient splitting, report external validation on data from other institutions and devices, and specify dataset demographics, like skin tone, anatomy site, and label origin [29], [70].

4. PREPROCESSING, SEGMENTATION, AUGMENTATION, AND CLASS IMBALANCE

4.1 Preprocessing

Preprocessing still needed in traditional pipelines was hair removal via morphological processing and inpainting, colour constancy, illumination correction, artefact removal [7], [8]. CNNs are generally more resilient to these variations, but in current DL papers most, if not all, of this excepting colour constancy, is either omitted or applied uniformly, usually with resizing and cropping and sometimes with intensity normalisation. Resolution is also critical: structures used in melanoma are often small, and a few top challenge solutions used multi-resolution inputs or large inputs [47]. Though, the impact of cropping and artefact removal must be taken into account: it can modify the information that networks have access to, leading to artefacts and shortcuts [31].

4.2 Lesion segmentation

Segmentation is used to determine the lesion border and has historically been the first step in CAD systems, as border- and shape-based features require it [6], [7]. Deep encoder-decoder structures, mainly U-Net [32] and its variants [33], are currently head-and-shoulders ahead of any competitors in the ISIC segmentation challenges. Employing transfer learning with encoders pretrained on classification improves segmentation accuracy. For classification, segmentation can be used to crop the lesion from the whole image to limit the spatial extent of the processing, as an auxiliary task for multi-task learning, or as a focus of the classifier. The merit of this approach varies: very deep residual models favor segmentation-

(This work is licensed under a Creative Commons Attribution 4.0 International License.)

International Journal of Engineering Research & Technology (IJERT)
ISSN: 2278-0181
Vol. 15 Issue 10 , October - 2026

aided classification [34], but many end-to-end classifiers give similar results to segmented systems, given sufficient training data, and errors in segmentation can propagate into the classifier.

4.3 Data augmentation

Due to the relative lack of annotated dermatology images, augmentation techniques are common. Rotation flipping scaling, cropping, and colour jittering are widely used, and a study of augmentation for skin lesion analysis found that a systematic approach to augmentation can produce gains comparable to model architecture changes [35]. Also, researchers can generate new lesions using generative adversarial networks or diffusion-based methods, though a survey of methods concluded that synthetic data often adds limited benefit, can leak or distort clinical features, and should be evaluated for realism and diversity [36]. But, synthetic images could be useful for anonymisation and rare classes if validated by clinicians.

4.4 Class imbalance

In HAM10000, melanocytic nevi account for roughly two thirds of images, while dermatofibroma and vascular lesions together account for only a few percent [22]. In SIIM-ISIC 2020, melanoma represents under 2% of the images [25]. Without mitigation, models optimise accuracy by favouring the majority class. Common remedies include re-sampling, weighted losses, focal loss, which down-weights easy examples [37], and class-balanced loss based on the effective number of samples [38]. Evaluation should use metrics beyond accuracy, such as per-class sensitivity and specificity, balanced accuracy, area under the ROC curve, and precision-recall measures. Threshold selection should reflect clinical priorities: missing a melanoma is generally costlier than an unnecessary referral, but a high false-positive rate overloads clinics and biopsy services.

5. CONVOLUTIONAL NEURAL NETWORK APPROACHES

5.1 Transfer learning and backbone architectures

Because skin lesion data sets are much smaller than natural-image data sets, almost all of the recently published state-of-the-art algorithms were based on networks that had been pretrained on ImageNet [10], [11]. In the computer-vision community, backbone architectures have evolved from AlexNet [11], through Inception [39], residual networks (ResNet) [40], densely connected networks (DenseNet) [41], and now compound-scaled EfficientNets [42]. For instance, Esteva et al. fine-tuned an Inception-v3 network [12], Haenssle et al. used a dermoscopy-specific Inception variant [13], Yu et al. demonstrated the strength of deep residual networks in recognising melanoma [34], Han et al. trained a ResNet on clinical images of benign and malignant skin tumours [45], and Mahbod et al. showed how combining deep features from several fine-tuned networks leads to better classification [44].

5.2 Ensembles and multi-resolution models

Ensembling is the strongest way to improve challenge results. Codella et al. used deep residual features, sparse coding and segmentation metadata and showed all ensembles beat the accuracy of individual models in ISIC 2016 [43]. Gessert et al. used multi-resolution EfficientNet ensemble with metadata, with first place result in ISIC 2019 [47]. Kassem et al. achieved very high multiclass accuracy on ISIC 2019 using transfer learning with GoogleNet [46]. Ensembles also affect cost of computation and regulation validation and deployment on mobile hardware.

5.3 Multi-task and attribute-based learning

Because dermatologists reason through structured features like pigment networks, globules, streaks, and blue-white veils, researchers have used these attributes as auxiliary supervision. The seven-point checklist approach of Kawahara et al. jointly predicts diagnosis and dermoscopic criteria with a multi-task network [56], improving both performance and interpretability. Such methods link learned representations with clinical vocabulary and are an appealing route to explainable systems.

Table 3. Representative deep learning studies on skin lesion classification.

Study	Task / data	Model	Key finding

(This work is licensed under a Creative Commons Attribution 4.0 International License.)

International Journal of Engineering Research & Technology (IJERT)
ISSN: 2278-0181
Vol. 15 Issue 10 , October - 2026

Study	Task / data	Model	Key finding
Esteva et al. [12]	Clinical images (129,450); 2,032 diseases	Inception-v3, transfer learning	Dermatologist-level performance versus 21 dermatologists on selected tasks
Haenssle et al. [13]	Dermoscopic melanoma vs. nevus	Inception-based CNN	Outperformed most of 58 dermatologists
Yu et al. [34]	ISIC 2016	Very deep residual networks	Segmentation-assisted melanoma recognition
Han et al. [45]	Clinical images, cutaneous tumours	ResNet	Classified 12 skin diseases; external testing performed
Codella et al. [43]	ISIC 2016	Ensemble of deep and classical features	Ensembling improved over single models
Gessert et al. [47]	ISIC 2019, eight classes	Multi-resolution EfficientNet ensemble + metadata	Challenge-winning approach
Liu et al. [58]	26 common skin conditions, clinical images + history	Deep learning system	Accuracy comparable to dermatologists, above primary care physicians and nurse practitioners
Pacheco and Krohling [57]	PAD-UFES-20	CNN + attention-based metadata fusion	Metadata improved smartphone-image classification

5.4. Critical appraisal of the evidence base

Though the representative studies in Table 3 show progress, they should not be ranked as if they came from the same experiment. Certain choices are made differently across studies for example modality diagnostic task, class prevalence, reference standard, reader expertise, threshold, and whether the interpretation was internal or external. The table below indicates the minimum checks on interpretation that should be made in support of a performance claim. Actual values should only be copied from the source articles after checking.

Table 3A. Evidence-interpretation checks for representative studies.

Study group	Task / evidence setting	Strength	Interpretation checks before citing performance
Esteva et al. [12]	Clinical-image classification across selected disease categories	Reader-comparison design and large training collection are notable strengths.	Verify the exact test-set composition, task-specific sensitivity/specificity, confidence intervals, and whether the comparator task matches the intended use.
Haenssle et al. [13]	Dermoscopy; melanoma versus nevus	Direct comparison with dermatologists makes the result clinically interpretable for the tested task.	Do not generalise the result to all skin lesions, populations, devices, or routine workflows; verify reader conditions and operating threshold.
Brinker et al. [63], [64]; Maron et al. [65]	Reader studies across dermoscopic or clinical image tasks	Provide comparative evidence beyond model-only benchmark scores.	Reader expertise, case spectrum, image selection, and binary versus multiclass endpoints differ; report these before comparing results.
Tschandl et al. [67], [68]	Algorithm-versus-reader and human–AI collaboration studies	Highlight both standalone algorithm performance and the role of decision support.	Distinguish retrospective image classification from prospective clinical benefit; evaluate automation bias and subgroup performance.
Challenge and benchmark studies [20], [21], [23], [47]	Curated challenge datasets and held-out benchmark splits	Support reproducible within-benchmark comparison.	Challenge rankings do not establish external validity; check patient/lesion-level splitting, duplicate control, data provenance, and external testing.

(This work is licensed under a Creative Commons Attribution 4.0 International License.)

International Journal of Engineering Research & Technology (IJERT)
ISSN: 2278-0181
Vol. 15 Issue 10 , October - 2026

Multimodal studies [47], [57], [58]	Image data combined with metadata, multiple views, or clinical history	Better reflect the information available in clinical assessment.	Check missing-data handling, metadata leakage, ablation results, and performance when a modality is unavailable.

6. TRANSFORMERS, SELF-SUPERVISED LEARNING, AND FOUNDATION MODELS

6.1 Vision transformers

Self-attention [48] based transformer architecture was transferred on to images using splitting images into patches and treating them as tokens sequence (Vision Transformer, ViT) [49]. Hierarchical type variants such as the Swin Transformer introduced shifted windows efficiently modeling local-global interactions [50]. In skin lesions, these transformations captured the long-range interactions between lesion areas, and after pretraining on large data sets, they have matched or even exceeded the CNNs in skin imaging [52]. Matsoukas et al. addressed the issue of whether transformers should displace CNNs in all medical images, and proved that with pretraining, they can match CNNs, but their performance strongly depends on the pretraining method and data size [51]. The hybrid combination of CNNs with transformer models or two family solutions averaging, is the best compromise today, since CNNs implicitly include a local prior that is extremely relevant under-sampled conditions.

6.2 Self-supervised and contrastive learning

Labelled dermatology images are expensive to obtain, while unlabelled clinical photographs are widely available. Contrastive self-supervised methods like SimCLR learn image representations by comparing different augmented versions of the same image [53]. Azizi et al. found that self-supervised pretrainingincluding a multi-instance contrastive approach using multiple views from the same patientimproved accuracy and label efficiency for both dermatology condition classification and chest radiograph tasks, and increased robustness to distribution shifts [54]. These findings indicate that self-supervision is well suited to dermatology, where multiple images per lesion or patient are often accessible.

6.3 Foundation and multimodal models

The concept of foundation models [55], originally developed for generalisable learning from data and generalisable to numerous domains, is emerging in medicine [55]. They are promising for generalist applications with flexible zero- or few-shot classification, report generation, and image-text reasoning. In dermatology, early reports show promising data efficiency and generalisability across sites, but testing across skin tones and rare conditions and real-world clinical workflows remains elusive; questions about hallucinating or unsupported explanations, opaque training data, and regulatory validation for open-ended use are raised.

7. MULTIMODAL FUSION OF IMAGES AND PATIENT DATA

Dermatologists base diagnosis on multiple factors. Features like age sex area, lesion history, symptoms (e.g. bleeding or itching), as well as patient and family history of skin cancer, influence the picture diagnosis. Several studies combined images with tabular meta-data. Simple concatenation of meta-data features with CNN embeddings was common and was used by the top challenge entries [47]. More sophisticated attention-based meta-data image fusion, which allows meta-data to influence image features improved smartphone-image classification on PAD-UFES-20 [57]. Liu et al. created a multi-image multi-source medical information system, which predicts 26 conditions at expert level [58]. Approaches like lesion comparison with other nevi (the "ugly duckling" sign), and multi-image diagnosis are promising, but rely on patient-level datasets like SIIM-ISIC 2020 retaining patient identifiers [25].

There are additional risks of fusion. Models can become too dependent on "metadatum shortcuts" (e.g. location of a sample in the body is strongly linked to its diagnosis in the training data), and similar shortcuts at deployment time and

(This work is licensed under a Creative Commons Attribution 4.0 International License.)

International Journal of Engineering Research & Technology (IJERT)
ISSN: 2278-0181
Vol. 15 Issue 10 , October - 2026

missing/corrupted data can hurt accuracy. Development of robust systems should involve training and testing with realistic missingness and measuring the effect of adding in metadata

.8. EXPLAINABILITY, INTERPRETABILITY, AND UNCERTAINTY

Ad hoc visual explanation methods like Grad-CAM, which highlights the image regions most influencing a prediction [59], and model-agnostic techniques like LIME [60] are commonly used in skin lesion research. These methods assist with debugging, for example by revealing when a model focuses on rulers, ink markings, or image borders instead of the lesion itself [31]. But, they have recognized limitations: saliency maps can be unstable, may not represent the model's true reasoning, and might provide clinicians with false reassurance. Rudin has argued that for high-stakes decisions, inherently interpretable models are preferable to explanations of black-box models [61]. In dermatology, this perspective supports concept-based and attribute-based architectures that predict clinically relevant features before diagnosis [56], as well as case-based or prototype reasoning that presents similar reference lesions.

Uncertainty estimation is also important. A system that can abstain, defer, or flag out-of-distribution inputs is safer than one that always brings a confident prediction for every image. Monte Carlo dropout treats dropout during inference as approximate Bayesian inference [62]. Other practical methods include deep ensembles, test-time augmentation, and conformal prediction. The ISIC 2019 challenge included an out-of-distribution class to assess whether models can identify lesion types not seen during training [23], and many models still struggle with this. Calibration where predicted probabilities match observed frequencies should be reported, as uncalibrated scores can mislead clinicians and affect downstream decision-making..

9. ALGORITHM VERSUS CLINICIAN EVIDENCE

There have been a handful of studies comparing CNNs to dermatologists since 2017 (Table 3). They include the study by Esteva et al. who demonstrated that their CNN was able to demonstrate an expertise equivalent to 21 board certified dermatologists when conducting binary classification tasks on biopsy-proven images [12], Hnssle et al. who demonstrated that the area under the receiver operating characteristic (ROC) curve for a CNN was higher than the average of 58 dermatologists when dermoscopic recognition of melanoma, with little enhancement when the clinical context was provided [13], Brinker et al. who demonstrated the ability for a CNN to outperform 136 of 157 dermatologists when discriminating melanoma from nevi in dermoscopic images and perform to the same level as 145 dermatologists when presented with clinical images, Maron et al. who demonstrated the ability of a CNN to outperform 112 dermatologists when performing a five-class dermoscopic classification [65], and the best algorithms from the ISIC 2016 challenge analysis, which outperformed the average reader but not the best individual dermatologists [66] [63] [64].

Table 3. Selected algorithm versus clinician comparisons.

Study	Readers	Image type	Main finding
Esteva et al. [12]	21 dermatologists	Clinical / dermoscopic, biopsy-proven	CNN on par with dermatologists
Haenssle et al. [13]	58 dermatologists	Dermoscopy	CNN outperformed most readers
Brinker et al. [63]	157 dermatologists	Dermoscopy	CNN outperformed 136 of 157
Brinker et al. [64]	145 dermatologists	Clinical images	CNN on par with readers
Maron et al. [65]	112 dermatologists	Dermoscopy, five classes	CNN outperformed readers
Marchetti et al. [66]	8 dermatologists (ISIC 2016 test comparison)	Dermoscopy	Top algorithms outperformed the average reader

(This work is licensed under a Creative Commons Attribution 4.0 International License.)

International Journal of Engineering Research & Technology (IJERT)
ISSN: 2278-0181
Vol. 15 Issue 10 , October - 2026

Study	Readers	Image type	Main finding
Tschandl et al. [67]	511 readers vs. 139 algorithms	Dermoscopy, seven classes	Top algorithms better on average; humans more robust on out-of-distribution images
Tschandl et al. [68]	Clinicians with AI support	Dermoscopy	AI assistance helped, with benefit depending on the type of AI interface and reader experience

While these results are promising, they should be viewed with a level of caution. Reader studies are generally retrospective, use pre-selected images, force binary decisions without the context of a patient, and may not simulate the true prevalence of cases. A large multi-national study where 511 human readers competed against 139 algorithms on a 7-class task reported similar results: the best algorithms outperformed the average reader overall, though readers performed better on images that differ from the training set [67]. A meta-analysis of CAD systems for melanoma in the same way reported promising accuracy, but high levels of bias and mix across the studies [16].

Perhaps the most practically significant question is whether AI makes clinicians better rather than replaces them. Tschandl et al. demonstrated that AI support improved the accuracy of individual reader, but that this effect depended on the kind of AI support, the experience of the reader, and that some readers could be led astray by wrong AI suggestions [68]. This is a good example of the importance of research into human-AI interactions, automation bias, and workflow design that complements stand alone algorithm performance metrics.

10. CHALLENGES AND OPEN PROBLEMS

10.1 Dataset bias and skin-tone fairness

Disparities in performance for different skin tones are among the most pressing issues. Adamson and Smith warned: 'the use of such non-diverse training data might make worse disparities in health outcomes for more diverse populations' [69]. Their concern has since been validated: the DDI study found that state-of-the-art dermatology algorithms were much less accurate on darker skin tones and rare diseases, and that fine-tuning on balanced data could reduce, but not eliminate, the gap [28]. Groh et al. noted that models trained on clinical images perform best on skin types similar to their training data [27], while Kinyanjui et al. demonstrated that skin-tone estimation is unreliable and that its effect on performance is hard to disentangle from other factors [30]. A survey of dermatology datasets revealed frequent lack of information on skin-tone, ethnicity, and location [29]. These problems can be addressed by targeted collection of diverse data, explicit skin-tone annotation, and fairness-aware testing broken down by subgroup, rather than reporting only average accuracy.

10.2 Shortcut learning and robustness

All artefacts that help reveal the label in the training set can be exploited by models. Known artefacts are surgical markings [31] rulers colour patches, and image-source signatures. Evaluation of skin lesion AI in the face of real-world input changes, like contrast focus rotation, and lighting, uncovered deficits in clinic readiness [71]. Adversarial examples are effective against medical image classifiers, which could be exploited in the presence of financial motivation [72]. Robustness should be tested with external validation, perturbation testing, and in the presence of data from other sources and environments.

10.3 Domain shift and generalisation

Dermoscopes, cameras, smartphones, and acquisition protocols vary across institutions. Models trained on a single dermoscopic archive tend to deteriorate when applied to clinical or smartphone images [26], [27]. Domain adaptation, test-time adaptation, heavy augmentation, and self-supervised pretraining on different datasets [54] could work, but few methods report truly external or prospective validation. The performance of algorithm-based smartphone apps shows this

(This work is licensed under a Creative Commons Attribution 4.0 International License.)

International Journal of Engineering Research & Technology (IJERT)
ISSN: 2278-0181
Vol. 15 Issue 10 , October - 2026

gap: a systematic review identified limited and poor quality evidence on their performance, and the risk that they could miss melanomas [18].

10.4 Reporting standards and reproducibility

Although thousands of studies are published in the field, critical details that allow replication or assessment are regularly missing: information on patient-level splitting, the inclusion or exclusion of duplicate images, label sources, class distributions, and patient demographics. Consensus guidelines for reporting image-based AI studies in dermatology (the CLEAR Derm guidelines) has a checklist for data, model development, validation, and reporting of skin-tone and clinical context [70]. Applying standards like this would improve comparability and confidence.

10.5 Privacy, data sharing, and federated learning

Identifiable images of skin (for example the face) and data-sharing agreements preclude aggregation of data from different institutions. Federated learningin which multiple institutions participate in the training of a common model without sharing data [73]may enable access to larger, more diverse training data while maintaining privacy, and has been proposed as a foundation for future digital health [74]. Obstacles include system mix, communication costs, privacy attacks and challenges to auditing non-local data quality.

10.6 Clinical integration and regulation

The accuracy alone is not sufficient to warrant clinical utility; workflow integration is the other crucial step. For some examples, how do you define the implied use triaging, second reading, or autonomous diagnosis; who is legally or ethically responsible if the algorithm errs; how do we track for performance drift after the algorithm is in the field; does it improve patient outcomes, such as stage at presentation or a reduction in biopsies; does the device lead to over-diagnosis of indolent lesions, in particular if algorithms pick up benign lesions? There are few prospective trials in naturalistic pathways compared to landmark retrospective studies.

10.7. Limitations of this review

Limitations and interpretations of this review this review has these limitations. Its narrative nature does not allow reproducible systematic search or formal risk-of-bias assessment. Its scope may so be incomplete. Its various studies use heterogeneous datasets, image ways, class definitions, reference standards and evaluation protocols, precluding direct ranking of reported performance, or a defensible pooled estimate. Some rapidly changing subfields, Worth noting foundation and vision-language models, may have changed since the cited studies. As is the case with all back-of-the-envelope benchmarks and reader-studies, benchmark, and reader-study results alone do not establish real-world clinical benefits or safe implementation. These limitations lead us to interpret this review first and foremost as a qualitative map of the field, rather than a formalised exercise in evidence grading.

11. FUTURE RESEARCH DIRECTIONS

1. Detailed patient level information. Collecting data from many sites, several ethnic groups, biopsy-proven condition, skin-tone labels, follow-up, and total-body photography would limit bias and also enable patient-specific inference [25], [28], [29].

2. Prospective and external validation Multi-centre prospective studies in intended use settings with prespecified end points and predefiend subgroups are required to move forward from benchmarking with retrospective data [16], [70].

3. In particular, foundation, and self-supervised models, with dedicated intensive pretraining on a large amount of unlabelled and multimodal dermatology data, could boost label efficiency and robustness with an exhaustive audit for bias and hallucination [54], [55].

4. Decision-support interfaces. We should investigate how decision-support interfaces affect reader accuracy, trust, and automation bias [68].

(This work is licensed under a Creative Commons Attribution 4.0 International License.)

International Journal of Engineering Research & Technology (IJERT)
ISSN: 2278-0181
Vol. 15 Issue 10 , October - 2026

5. Interpretable and concept-based models. Decisions can be made auditable by predicting clinically relevant attributes and showing reference cases [56], [61].

6. Uncertainty, abstention, and out-of-distribution detection Well-calibrated models abstaining from difficult cases are essential for safe deployment [23], [62].

7. Training without disclosing sensitive data there is no requirement to centralise the images [73], [74] using federated and distributed training,.

8. Fairness-aware training and evaluation Fairness-aware training and evaluation methods for sub-group bias mitigation (e.g. reweighting, targeted collection, adversarial debiasing) should be evaluated using standardised sub-group reporting [27], [28], [30].

9. Beyond melanoma, most of the studies addressed the melanoma and some additional common groups; differential diagnosis, inflammatory/infectious diseases and other rare conditions deserve similar attention [58].

11.1. A practical clinical-readiness framework

A model should not be considered deployment-ready on the basis of a high internal test score alone. The following checklist can be used to structure future studies and to make the evidence needed for clinical translation explicit.

Table 4. Minimum domains for evaluating clinical readiness.

Domain	Minimum reporting / evaluation
Data provenance and representativeness	Report institutions, devices, geography, skin-tone representation, age range, lesion sites, class prevalence, label source, and missingness.
Leakage-resistant evaluation	Split at patient and, where relevant, lesion level; disclose duplicate handling and keep external data separate from model development.
Diagnostic performance	Report sensitivity, specificity, ROC-AUC, precision–recall AUC where imbalance is substantial, macro-F1/balanced accuracy for multiclass tasks, confidence intervals, and clinically relevant thresholds.
Calibration and uncertainty	Assess calibration, selective prediction/abstention, out-of-distribution detection, and the consequences of false negatives and false positives.
External and prospective validation	Test across independent institutions, devices, and populations; where feasible, prospectively evaluate workflow and patient-relevant endpoints.
Fairness and subgroup analysis	Report performance and uncertainty across skin tone, age, sex, anatomical site, and other clinically relevant groups; state when subgroup data are insufficient.
Clinical utility and human factors	Compare with appropriate clinicians and evaluate AI-assisted workflows, automation bias, referral/biopsy consequences, and usability.
Reproducibility and governance	Provide model/data provenance, preprocessing, threshold selection, code or implementation details where permissible, privacy safeguards, intended use, and post-deployment monitoring plans.

12. CONCLUSION

Deep learning has revolutionized skin lesion classification. Since the initial demonstrations of dermatologist-level performance [12], [13] and the introduction of larger challenges, multi-resolution ensembles [47], transformers [49], [50], self-supervised learning [54], and multimodal systems [57], [58] technical advances have been significant. Reader studies demonstrate that algorithms can now be competitive or outperform clinicians on curated challenges [63][65] and the greatest potential for benefit may come through humanAI partnerships [68]. Yet success on benchmark challenges does not necessarily imply clinical utility. Challenges such as dataset bias and under-representation of darker skin types [28], [29], shortcut learning [31], domain shift [71], inconsistent reporting [70], and limited prospective evidence [16], [18] remain. Future efforts should because of this prioritise improving the representativeness robustness transparency, and evidence of clinical benefit of systems, rather than focusing on incremental improvements in benchmark accuracy. Delivering on

(This work is licensed under a Creative Commons Attribution 4.0 International License.)

International Journal of Engineering Research & Technology (IJERT)
ISSN: 2278-0181
Vol. 15 Issue 10 , October - 2026

progress in these dimensions will determine whether AI-assisted skin cancer classification can become a trusted part of dermatological practice around the world.

REFERENCES

[1] H. Sung, J. Ferlay, R. L. Siegel, et al., "Global cancer statistics 2020: GLOBOCAN estimates of incidence and mortality worldwide for 36 cancers in 185 countries," *CA Cancer J. Clin.*, vol. 71, no. 3, pp. 209–249, 2021.
[2] R. L. Siegel, A. N. Giaquinto, and A. Jemal, "Cancer statistics, 2024," *CA Cancer J. Clin.*, vol. 74, no. 1, pp. 12–49, 2024.
[3] H. Kittler, H. Pehamberger, K. Wolff, and M. Binder, "Diagnostic accuracy of dermoscopy," *Lancet Oncol.*, vol. 3, no. 3, pp. 159–165, 2002.
[4] N. R. Abbasi, H. M. Shaw, D. S. Rigel, et al., "Early diagnosis of cutaneous melanoma: revisiting the ABCD criteria," *JAMA*, vol. 292, no. 22, pp. 2771–2776, 2004.
[5] G. Argenziano, G. Soyer, V. De Giorgio, et al., "Dermoscopy of pigmented skin lesions: results of a consensus meeting via the Internet," *J. Am. Acad. Dermatol.*, vol. 48, no. 5, pp. 679–693, 2003.
[6] H. Ganster, A. Pinz, R. Röhrer, et al., "Automated melanoma recognition," *IEEE Trans. Med. Imaging*, vol. 20, no. 3, pp. 233–239, 2001.
[7] M. E. Celebi, H. A. Kingravi, B. Uddin, et al., "A methodological approach to the classification of dermoscopy images," *Comput. Med. Imaging Graph.*, vol. 31, no. 6, pp. 362–373, 2007.
[8] C. Barata, M. E. Celebi, and J. S. Marques, "A survey of feature extraction in dermoscopy image analysis of skin cancer," *IEEE J. Biomed. Health Inform.*, vol. 23, no. 3, pp. 1096–1109, 2019.
[9] Y. LeCun, Y. Bengio, and G. Hinton, "Deep learning," *Nature*, vol. 521, no. 7553, pp. 436–444, 2015.
[10] G. Litjens, T. Kooi, B. E. Bejnordi, et al., "A survey on deep learning in medical image analysis," *Med. Image Anal.*, vol. 42, pp. 60–88, 2017.
[11] A. Krizhevsky, I. Sutskever, and G. E. Hinton, "ImageNet classification with deep convolutional neural networks," in *Proc. Adv. Neural Inf. Process. Syst. (NIPS)*, 2012, pp. 1097–1105.
[12] 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.
[13] H. A. Haenssle, C. Fink, R. Schneiderbauer, et al., "Man against machine: diagnostic performance of a deep learning convolutional neural network for dermoscopic melanoma recognition in comparison to 58 dermatologists," *Ann. Oncol.*, vol. 29, no. 8, pp. 1836–1842, 2018.
[14] A. Takiddin, J. Schneider, Y. Yang, A. Abd-Alrazaq, and M. Househ, "Artificial intelligence for skin cancer detection: scoping review," *J. Med. Internet Res.*, vol. 23, no. 11, e22934, 2021.
[15] Y. Wu, B. Chen, A. Zeng, D. Pan, R. Wang, and S. Zhao, "Skin cancer classification with deep learning: a systematic review," *Front. Oncol.*, vol. 12, 893972, 2022.
[16] V. Dick, C. Sinz, M. Mittlböck, H. Kittler, and P. Tschandl, "Accuracy of computer-aided diagnosis of melanoma: a meta-analysis," *JAMA Dermatol.*, vol. 155, no. 11, pp. 1291–1299, 2019.
[17] A. Hekler, J. S. Utikal, A. H. Enk, et al., "Deep learning outperformed 11 pathologists in the classification of histopathological melanoma images," *Eur. J. Cancer*, vol. 118, pp. 91–96, 2019.
[18] K. Freeman, J. Dinnes, N. Chuchu, et al., "Algorithm based smartphone apps to assess risk of skin cancer in adults: systematic review of diagnostic accuracy studies," *BMJ*, vol. 368, m127, 2020.
[19] T. Mendonça, P. M. Ferreira, J. S. Marques, A. R. S. Marcal, and J. Rozeira, "PH2 – a dermoscopic image database for research and benchmarking," in *Proc. 35th Annu. Int. Conf. IEEE Eng. Med. Biol. Soc. (EMBC)*, 2013, pp. 5437–5440.
[20] D. Gutman, N. C. F. Codella, E. Celebi, et al., "Skin lesion analysis toward melanoma detection: a challenge at the International Symposium on Biomedical Imaging (ISBI) 2016, hosted by the International Skin Imaging Collaboration (ISIC)," arXiv:1605.01397, 2016.
[21] 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 (ISBI), hosted by the International Skin Imaging Collaboration (ISIC)," in *Proc. IEEE 15th Int. Symp. Biomed. Imaging (ISBI)*, 2018, pp. 168–172.
[22] P. Tschandl, C. Rosendahl, and H. Kittler, "The HAM10000 dataset, a large collection of multi-source dermatoscopic images of common pigmented skin lesions," *Sci. Data*, vol. 5, 180161, 2018.
[23] N. Codella, V. Rotemberg, P. Tschandl, et al., "Skin lesion analysis toward melanoma detection 2018: a challenge hosted by the International Skin Imaging Collaboration (ISIC)," arXiv:1902.03368, 2019.
[24] M. Combalia, N. C. F. Codella, V. Rotemberg, et al., "BCN20000: dermoscopic lesions in the wild," arXiv:1908.02288, 2019.
[25] V. Rotemberg, N. Kurtansky, B. Betz-Stablein, et al., "A patient-centric dataset of images and metadata for identifying melanomas using clinical context," *Sci. Data*, vol. 8, 34, 2021.
[26] 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 Brief*, vol. 32, 106221, 2020.
[27] M. Groh, C. Harris, L. Soenksen, et al., "Evaluating deep neural networks trained on clinical images in dermatology with the Fitzpatrick 17k dataset," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. Workshops (CVPRW)*, 2021, pp. 1820–1828.
[28] R. Daneshjou, K. Vodrahalli, R. A. Novoa, et al., "Disparities in dermatology AI performance on a diverse, curated clinical image set," *Sci. Adv.*, vol. 8, no. 32, eabq6147, 2022.
[29] R. Daneshjou, M. P. Smith, M. D. Sun, V. Rotemberg, and J. Zou, "Lack of transparency and potential bias in artificial intelligence data sets and algorithms: a scoping review," *JAMA Dermatol.*, vol. 157, no. 11, pp. 1362–1369, 2021.
[30] N. M. Kinyanjui, T. Odonga, C. Cintas, et al., "Estimating skin tone and effects on classification performance in dermatology datasets," arXiv:1910.13268, 2019.
[31] J. K. Winkler, C. Fink, F. Toberer, et al., "Association between surgical skin markings in dermoscopic images and diagnostic performance of a deep learning convolutional neural network for melanoma recognition," *JAMA Dermatol.*, vol. 155, no. 10, pp. 1135–1141, 2019.
[32] O. Ronneberger, P. Fischer, and T. Brox, "U-Net: convolutional networks for biomedical image segmentation," in *Proc. Int. Conf. Med. Image Comput. Comput.-Assist. Interv. (MICCAI)*, 2015, pp. 234–241.
[33] P. Tschandl, C. Sinz, and H. Kittler, "Domain-specific classification-pretrained fully convolutional network encoders for skin lesion segmentation," *Comput. Biol. Med.*, vol. 104, pp. 111–116, 2019.

(This work is licensed under a Creative Commons Attribution 4.0 International License.)

International Journal of Engineering Research & Technology (IJERT)
ISSN: 2278-0181
Vol. 15 Issue 10 , October - 2026

[34] L. Yu, H. Chen, Q. Dou, J. Qin, and P.-A. Heng, "Automated melanoma recognition in dermoscopy images via very deep residual networks," *IEEE Trans. Med. Imaging*, vol. 36, no. 4, pp. 994–1004, 2017.
[35] F. Perez, C. Vasconcelos, S. Avila, and E. Valle, "Data augmentation for skin lesion analysis," in *OR 2.0 Context-Aware Operating Theaters, Computer Assisted Robotic Endoscopy, Clinical Image-Based Procedures, and Skin Image Analysis*, Springer, 2018, pp. 303–311.
[36] A. Bissoto, E. Valle, and S. Avila, "GAN-based data augmentation and anonymization for skin-lesion analysis: a critical review," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. Workshops (CVPRW)*, 2021, pp. 1847–1856.
[37] T.-Y. Lin, P. Goyal, R. Girshick, K. He, and P. Dollár, "Focal loss for dense object detection," in *Proc. IEEE Int. Conf. Comput. Vis. (ICCV)*, 2017, pp. 2980–2988.
[38] Y. Cui, M. Jia, T.-Y. Lin, Y. Song, and S. Belongie, "Class-balanced loss based on effective number of samples," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, 2019, pp. 9268–9277.
[39] C. Szegedy, V. Vanhoucke, S. Ioffe, J. Shlens, and Z. Wojna, "Rethinking the Inception architecture for computer vision," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, 2016, pp. 2818–2826.
[40] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, 2016, pp. 770–778.
[41] G. Huang, Z. Liu, L. van der Maaten, and K. Q. Weinberger, "Densely connected convolutional networks," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, 2017, pp. 4700–4708.
[42] M. Tan and Q. V. Le, "EfficientNet: rethinking model scaling for convolutional neural networks," in *Proc. Int. Conf. Mach. Learn. (ICML)*, 2019, pp. 6105–6114.
[43] N. C. F. Codella, Q.-B. Nguyen, S. Pankanti, et al., "Deep learning ensembles for melanoma recognition in dermoscopy images," *IBM J. Res. Dev.*, vol. 61, no. 4/5, pp. 5:1–5:15, 2017.
[44] A. Mahbod, G. Schaefer, C. Wang, R. Ecker, and I. Ellinger, "Skin lesion classification using hybrid deep neural networks," in *Proc. IEEE Int. Conf. Acoust., Speech Signal Process. (ICASSP)*, 2019, pp. 1229–1233. (See also A. Mahbod et al., "Fusing fine-tuned deep features for skin lesion classification," *Comput. Med. Imaging Graph.*, vol. 71, pp. 19–29, 2019.)
[45] S. S. Han, M. S. Kim, W. Lim, G. H. Park, I. Park, and S. E. Chang, "Classification of the clinical images for benign and malignant cutaneous tumors using a deep learning algorithm," *J. Invest. Dermatol.*, vol. 138, no. 7, pp. 1529–1538, 2018.
[46] M. A. Kassem, K. M. Hosny, and M. M. Fouad, "Skin lesions classification into eight classes for ISIC 2019 using deep convolutional neural network and transfer learning," *IEEE Access*, vol. 8, pp. 114822–114832, 2020.
[47] N. Gessert, M. Nielsen, M. Shaikh, R. Werner, and A. Schlaefer, "Skin lesion classification using ensembles of multi-resolution EfficientNets with meta data," *MethodsX*, vol. 7, 100864, 2020.
[48] A. Vaswani, N. Shazeer, N. Parmar, et al., "Attention is all you need," in *Proc. Adv. Neural Inf. Process. Syst. (NIPS)*, 2017, pp. 5998–6008.
[49] A. Dosovitskiy, L. Beyer, A. Kolesnikov, et al., "An image is worth 16x16 words: transformers for image recognition at scale," in *Proc. Int. Conf. Learn. Represent. (ICLR)*, 2021.
[50] Z. Liu, Y. Lin, Y. Cao, et al., "Swin Transformer: hierarchical vision transformer using shifted windows," in *Proc. IEEE/CVF Int. Conf. Comput. Vis. (ICCV)*, 2021, pp. 10012–10022.
[51] C. Matsoukas, J. F. Haslum, M. Söderberg, and K. Smith, "Is it time to replace CNNs with transformers for medical images?" arXiv:2108.09038, 2021.
[52] S. Aladhadh, M. Alsanea, M. Aloraini, T. Khan, S. Habib, and M. Islam, "An effective skin cancer classification mechanism via medical vision transformer," *Sensors*, vol. 22, no. 11, 4008, 2022.
[53] T. Chen, S. Kornblith, M. Norouzi, and G. Hinton, "A simple framework for contrastive learning of visual representations," in *Proc. Int. Conf. Mach. Learn. (ICML)*, 2020, pp. 1597–1607.
[54] S. Azizi, B. Mustafa, F. Ryan, et al., "Robust and data-efficient generalization of self-supervised machine learning for diagnostic imaging," *Nat. Biomed. Eng.*, vol. 7, pp. 756–779, 2023.
[55] M. Moor, O. Banerjee, Z. S. H. Abad, et al., "Foundation models for generalist medical artificial intelligence," *Nature*, vol. 616, no. 7956, pp. 259–265, 2023.
[56] J. Kawahara, S. Daneshvar, G. Argenziano, and G. Hamarneh, "Seven-point checklist and skin lesion classification using multitask multimodal neural nets," *IEEE J. Biomed. Health Inform.*, vol. 23, no. 2, pp. 538–546, 2019.
[57] A. G. C. Pacheco and R. A. Krohling, "An attention-based mechanism to combine images and metadata in deep learning models applied to skin cancer classification," *IEEE J. Biomed. Health Inform.*, vol. 25, no. 9, pp. 3554–3563, 2021.
[58] Y. Liu, A. Jain, C. Eng, et al., "A deep learning system for differential diagnosis of skin diseases," *Nat. Med.*, vol. 26, no. 6, pp. 900–908, 2020.
[59] 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," in *Proc. IEEE Int. Conf. Comput. Vis. (ICCV)*, 2017, pp. 618–626.
[60] M. T. Ribeiro, S. Singh, and C. Guestrin, ""Why should I trust you?": explaining the predictions of any classifier," in *Proc. 22nd ACM SIGKDD Int. Conf. Knowl. Discov. Data Min.*, 2016, pp. 1135–1144.
[61] C. Rudin, "Stop explaining black box machine learning models for high stakes decisions and use interpretable models instead," *Nat. Mach. Intell.*, vol. 1, no. 5, pp. 206–215, 2019.
[62] Y. Gal and Z. Ghahramani, "Dropout as a Bayesian approximation: representing model uncertainty in deep learning," in *Proc. Int. Conf. Mach. Learn. (ICML)*, 2016, pp. 1050–1059.
[63] T. J. Brinker, A. Hekler, A. H. Enk, et al., "Deep learning outperformed 136 of 157 dermatologists in a head-to-head dermoscopic melanoma image classification task," *Eur. J. Cancer*, vol. 113, pp. 47–54, 2019.
[64] T. J. Brinker, A. Hekler, A. H. Enk, et al., "A convolutional neural network trained with dermoscopic images performed on par with 145 dermatologists in a clinical melanoma image classification task," *Eur. J. Cancer*, vol. 111, pp. 148–154, 2019.
[65] R. C. Maron, M. Weichenthal, J. S. Utikal, et al., "Systematic outperformance of 112 dermatologists in multiclass skin cancer image classification by convolutional neural networks," *Eur. J. Cancer*, vol. 119, pp. 57–65, 2019.
[66] M. A. Marchetti, N. C. F. Codella, S. W. Dusza, et al., "Results of the 2016 International Skin Imaging Collaboration International Symposium on Biomedical Imaging challenge: comparison of the accuracy of computer algorithms to dermatologists for the diagnosis of melanoma from dermoscopic images," *J. Am. Acad. Dermatol.*, vol. 78, no. 2, pp. 270–277, 2018.
[67] P. Tschandl, N. Codella, B. N. Akay, et al., "Comparison of the accuracy of human readers versus machine-learning algorithms for pigmented skin lesion classification: an open, web-based, international, diagnostic study," *Lancet Oncol.*, vol. 20, no. 7, pp. 938–947, 2019.

(This work is licensed under a Creative Commons Attribution 4.0 International License.)

International Journal of Engineering Research & Technology (IJERT)
ISSN: 2278-0181
Vol. 15 Issue 10 , October - 2026

[68] P. Tschandl, C. Rinner, Z. Apalla, et al., "Human–computer collaboration for skin cancer recognition," *Nat. Med.*, vol. 26, no. 8, pp. 1229–1234, 2020.
[69] A. S. Adamson and A. Smith, "Machine learning and health care disparities in dermatology," *JAMA Dermatol.*, vol. 154, no. 11, pp. 1247–1248, 2018.
[70] R. Daneshjou, C. Barata, B. Betz-Stablein, et al., "Checklist for evaluation of image-based artificial intelligence reports in dermatology: CLEAR Derm consensus guidelines from the International Skin Imaging Collaboration Artificial Intelligence Working Group," *JAMA Dermatol.*, vol. 158, no. 1, pp. 90–96, 2022.
[71] A. T. Young, K. Fernandez, J. Pfau, et al., "Stress testing reveals gaps in clinic readiness of image-based diagnostic artificial intelligence models," *npj Digit. Med.*, vol. 4, 10, 2021.
[72] S. G. Finlayson, J. D. Bowers, J. Ito, J. L. Zittrain, A. L. Beam, and I. S. Kohane, "Adversarial attacks on medical machine learning," *Science*, vol. 363, no. 6433, pp. 1287–1289, 2019.
[73] B. McMahan, E. Moore, D. Ramage, S. Hampson, and B. A. y Arcas, "Communication-efficient learning of deep networks from decentralized data," in *Proc. 20th Int. Conf. Artif. Intell. Stat. (AISTATS)*, 2017, pp. 1273–1282.
[74] N. Rieke, J. Hancox, W. Li, et al., "The future of digital health with federated learning," *npj Digit. Med.*, vol. 3, 119, 2020.
[75] Choksi, Prashansa & Bambhaniya, Vipul & Shingadiya, Chetan. (2026). Deep Transfer Learning for Automated Cervical Cell-Type Classification in Pap-Smear Images: A Comparative and Explainable Study. 10.21203/rs.3.rs-11060794/v1.
[76] Choksi, Prashansa & Shingadiya, Chetan. (2025). Cervical Cancer Detection Through Lens: A Review of Image Segmentation and Recognition Approaches. 10.1007/978-981-96-4391-2_40.

(This work is licensed under a Creative Commons Attribution 4.0 International License.)