REVIEW ARTICLE

Software-Engineering Perspectives on Machine for Skin-Disease Classification

Moomna Nazir[1,*], Azka Ahsan[2], Rabia Khadim[2], Shakeel Abbas[3], Aown Muhammad[4] and Zain Sohail[2]

[1] Department of Computer Science, Govt. Post Graduate College for Women, Sahiwal 57040, Pakistan
[2] Department of Computer Science, COMSATS University Islamabad, Sahiwal Campus, Sahiwal 57040, Pakistan
[3] Department of Computer Science, Minhaj University Lahore, Lahore 54770, Pakistan
[4] Department of Computer Science, University of Engineering and Technology, Lahore 54770, Pakistan

Abstract

Skin-disease classification has evolved from simple image recognizers into software-driven pipelines that demand reliability, reproducibility, and ethical governance. While most AI reviews focus on algorithmic accuracy, few examine these systems through a software-engineering (SE) lens—essential for assessing pipeline modularity, version control, deployment readiness, and long-term maintainability, all critical for clinical integration. This review surveys literature from 2015 to early 2025, curating about 180 papers that link skin-disease classification with SE practices. It traces the shift from handcrafted feature-based classifiers to end-to-end convolutional, ensemble, and transformer architectures, alongside the engineering processes that support versioning, deployment, and monitoring. Benchmark datasets (PH^2, HAM10000, ISIC, etc.) have established reproducible evaluation protocols that underpin software verification. Emerging directions—self-supervised pretraining, multimodal fusion, human-AI collaboration—signal a move from model-centric to system-level integration. The analysis highlights not only accuracy and generalization but also SE quality attributes: scalability, maintainability, explainability, and fairness, which are indispensable for trustworthy adoption in diverse clinical workflows.

Keywords: software engineering, machine learning, deep learning, dermatology, computer-aided diagnosis, MLOps, fairness.

1 Introduction

This review has three parts. First examine machine learning and deep learning approaches for skin-disease classification through a software-engineering perspective. Second, analyze engineering principles such as modularity, reproducibility, and deployment readiness. These principles shape the development and reliability of dermatology AI systems. Third, identify persistent

Submitted: 20 October 2025
Accepted: 16 December 2025
Published: 11 February 2026

Vol. 2, **No.** 1, 2026.
10.62762/JSE.2025.913699

***Corresponding author:**
✉ Moomna Nazir
moomnanazir@gmail.com

Citation
Nazir, M., Ahsan, A., Khadim, R., Abbas, S., Muhammad, A., & Sohail, Z. (2026). Software-Engineering Perspectives on Machine for Skin-Disease Classification. *ICCK Journal of Software Engineering*, 2(1), 52–70.

engineering, ethical, and operational challenges that must be addressed for safe and scalable clinical adoption.

Skin diseases including melanoma, basal cell carcinoma, and other chronic dermatological disorders remain among the world's most prevalent health challenges, affecting millions annually [1, 2]. Melanoma, with its high metastatic potential, has been recognized by the World Health Organization as one of the most aggressive malignancies [3]. Early and precise diagnosis is therefore critical for improving patient survival and reducing the cost of long-term care [4]. Traditional diagnosis depends on a clinician's visual inspection of dermoscopic or clinical images, a process that is inherently subjective and varies across observers. These limitations have motivated the development of software-driven computer-aided diagnostic systems that aim to standardize, scale, and integrate dermatological assessment within clinical workflows.

Dermatological disorders encompass a wide range of benign and malignant skin lesions that differ in morphology, pigmentation, and clinical presentation. The most common diagnostic categories include melanocytic nevi, vascular lesions, dermatofibroma, keratoses, and skin carcinomas such as basal cell carcinoma and squamous cell carcinoma. Representative examples of these categories are shown in Figure 1, illustrating the clinical diversity and visual complexity that make automated classification a challenging yet essential task in computer-aided dermatology.

Early ML pipelines relied on handcrafted color, texture, and shape features coupled to SVM/RF classifiers; performance hinged on feature quality and consistent segmentation [5, 6]. Although promising, these pipelines were constrained by small datasets, limited generalization, and an absence of standardized software engineering practices such as data-versioning or reproducible preprocessing. The introduction of deep learning, particularly convolutional neural networks (CNNs), marked a decisive shift toward end-to-end trainable systems. Seminal studies demonstrated dermatologist-level accuracy in differentiating malignant from benign lesions [8, 9]. Public datasets released through the International Skin Imaging Collaboration (ISIC) provided annotated images and reference protocols that enabled replication, benchmarking, and open-source collaboration [10, 11]. Subsequent innovations—EfficientNet, Vision Transformers (ViTs), self-supervised pre-training, and multimodal fusion with patient metadata—further evolved dermatology AI into a software ecosystem rather than a collection of isolated models [12, 13]. These traditional pipelines provided the foundational structure upon which modern, software-centric dermatology AI systems were built. Their limitations directly motivated the development of more modular, scalable, and reproducible deep-learning architectures. Moreover, these early pipelines lacked basic software-engineering practices such as version control and modular component design. This makes it difficult to update, debug, or maintain as new datasets and diagnostic requirements emerge.

Despite impressive progress, significant engineering and ethical challenges persist: severe class imbalance, under-representation of darker skin tones, domain-shift across imaging devices, and limited transparency in model behavior [13]. Recent frameworks such as CLEAR Derm have sought to codify reporting standards and promote reproducibility, fairness, and auditability in dermatology AI [14]. However, many studies still lack complete documentation of training environments, hyper-parameters, and validation of pipelines—issues that directly impede software reliability and regulatory compliance.

Applying a software-engineering lens in dermatology AI involves both model architectures and the entire pipeline as a software system. This includes examining data versioning, modular pipeline design, documentation quality, testing and validation practices, deployment workflows, and post-deployment monitoring. Such SE principles are essential because clinical AI must be reliable, traceable, maintainable, and safe qualities that arise from sound engineering rather than from model accuracy alone.

Despite extensive progress in algorithmic performance, existing reviews seldom address how these models translate into reproducible, maintainable, and ethically governed software systems. The intersection of dermatology AI and software engineering remains underexplored, particularly regarding data versioning, deployment workflows, and post-deployment monitoring. This paper bridges that gap by systematically analyzing ML and DL approaches through a software-engineering lens, emphasizing scalability, fairness, and operational reliability.

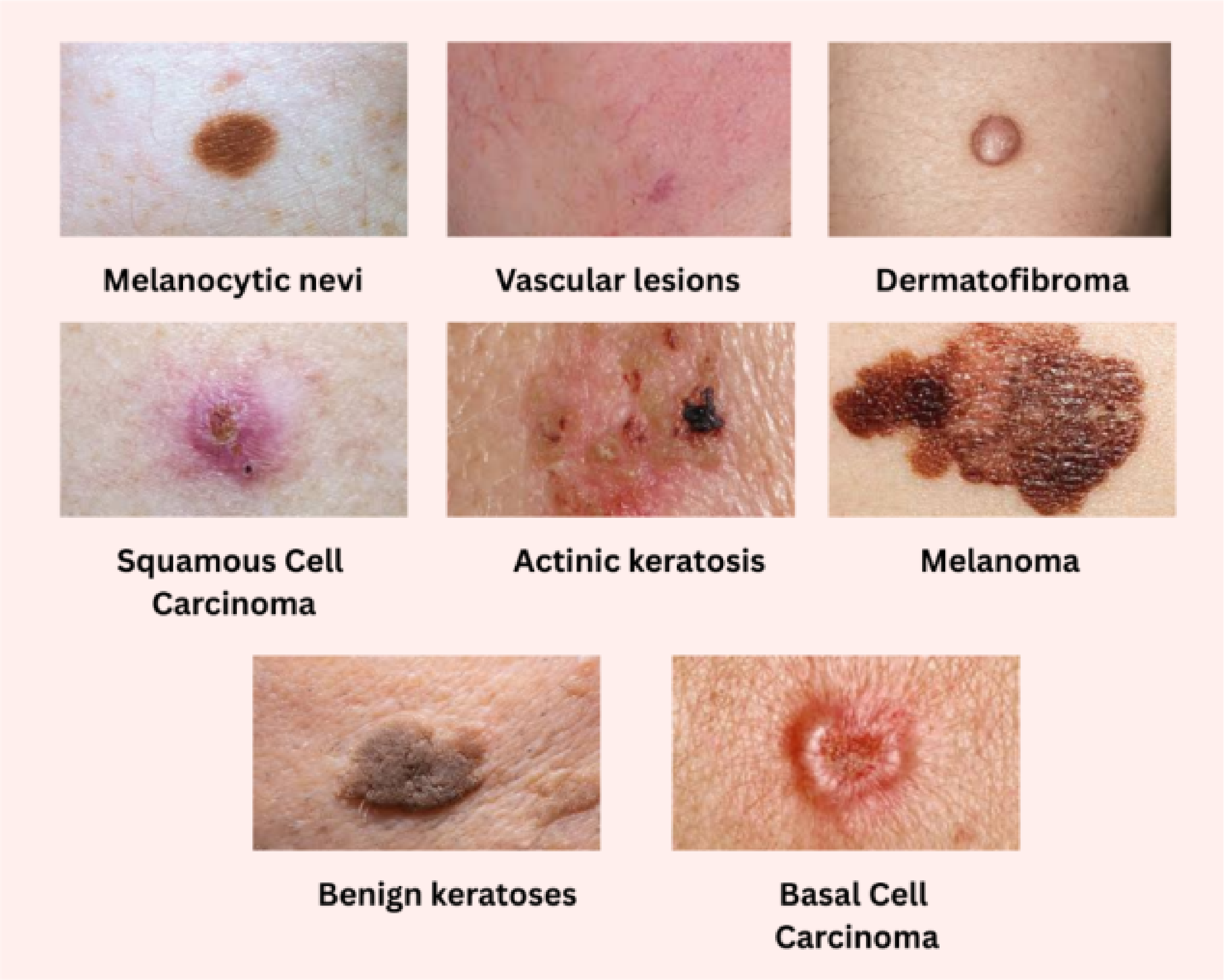

Figure 1. Representative categories of skin diseases commonly analyzed in dermatological image-classification research, illustrating the diversity of lesion types and diagnostic complexity encountered in automated analysis.

This review focuses on machine- and deep-learning–based skin-disease classification from 2015 to 2025 through a software engineering lens. This decade is particularly important because it captures the transition from handcrafted, non-standardized pipelines to modern deep-learning and transformer-based systems. This emergence of reproducibility standards, version control practices, dataset governance, and deployment-focused engineering in clinical AI. It traces the evolution from feature-engineered classifiers to modern deep architectures—CNNs, Transformers, multimodal, and self-supervised frameworks—while examining how reproducibility, maintainability, and fairness have been addressed. The scope includes common dermatologic conditions (melanoma, basal-cell, and squamous-cell carcinomas) as well as emerging efforts to model rare diseases and diverse skin tones. By connecting algorithmic progress with software-system considerations such as architecture design, version control, testing, and deployment, the paper highlights both scientific and engineering maturity in dermatology AI.

Overall, this work bridges medical AI research and software-engineering practice. Its main contributions are:

1. Combined Classical & Modern Analysis: Revisits traditional feature-based methods and compares them with modern deep-learning approaches to show how early pipelines shaped today's modular and reproducible dermatology AI systems.

2. Comprehensive Survey (2015–2025): Summarize advances in machine and deep learning for dermatology across data, models, and evaluation

practices.

3. Benchmark Resources: Review public datasets (ISIC, HAM10000, PH², BCN20000) and analyze how open benchmarks foster verification and reuse.

4. Architectural Innovations: Discuss EfficientNet, ViTs, and multimodal frameworks in terms of their software architecture patterns (scalability, portability, and resource profiling).

5. Challenges and Quality Attributes: Identify ongoing issues—dataset imbalance, algorithmic bias, limited interpretability—and relate them to software quality factors such as reliability and explainability.

6. Future Outlook: Outline directions for federated, fairness-aware, and explainable systems grounded in robust MLOps and governance practices.

2 Review Methodology

We adopted a systematic review methodology to ensure to ensure comprehensive and balanced reporting of machine-learning and deep-learning research in skin-disease classification. The search strategy targeted reviewed literature published between 2015 and early 2025. This period reflects the transition from traditional pipelines to modern deep-learning and transformer-based dermatology of AI systems.

Search Sources: Relevant studies were searched through major scientific databases, including IEEE Xplore, PubMed, Scopus, ScienceDirect, and Google Scholar. These databases were selected to find publications spanning computer vision, medical imaging, software engineering, and clinical research.

Search Strategy: Search queries combined domain-specific and software-engineering-related terms, including "skin disease classification," "dermoscopy," "computer-aided diagnosis," "machine learning," "deep learning," "CNN," "Vision Transformer," "multimodal dermatology AI," "pipeline reproducibility," "software engineering," "MLOps," and "model deployment."

2.1 Inclusion Criteria

Studies were included if they met the following criteria:

1. Published between 2015 and early 2025 in peer-reviewed venues.

2. Proposed or evaluated ML/DL-based skin disease classification methods.

3. Reported methodological details, datasets, performance metrics, or software engineering considerations.

4. Used dermoscopic, clinical, or multimodal dermatology imaging data.

5. Demonstrated relevance to system design, reproducibility, deployment, or engineering practices.

2.2 Exclusion Criteria

The following were excluded:

1. Clinical papers without computational or system-relevant methods.

2. Articles lacking methodological transparency or reproducibility details.

3. Review papers that did not present a new analysis.

4. Non-English publications and preprints lacking peer review.

2.3 Study Selection Workflow

In the first search, almost 420 studies are derived. After removing duplicates, 358 articles remained for screening. Title and abstract review reduced this set to 226 full-text articles. Out of which approximately 180 met all inclusion criteria and were included in the final synthesis. A simplified selection workflow is presented in Figure 2.

3 Traditional approaches for skin disease detection

Before the deep learning revolution, classical computer-aided diagnosis (CAD) systems formed the foundation of automated skin disease analysis. These early pipelines followed a deterministic software architecture consisting of four sequential modules preprocessing, segmentation, feature extraction, and classification. In the preprocessing phase, image quality was enhanced and artifacts such as hair and noise were removed using algorithms like DullRazor, Gaussian filtering, and color normalization. Segmentation techniques including thresholding, region growing, watershed, and active contours were then applied to delineate lesions from surrounding tissue [5, 15].

A crucial component of these systems was handcrafted feature engineering, where domain knowledge guided

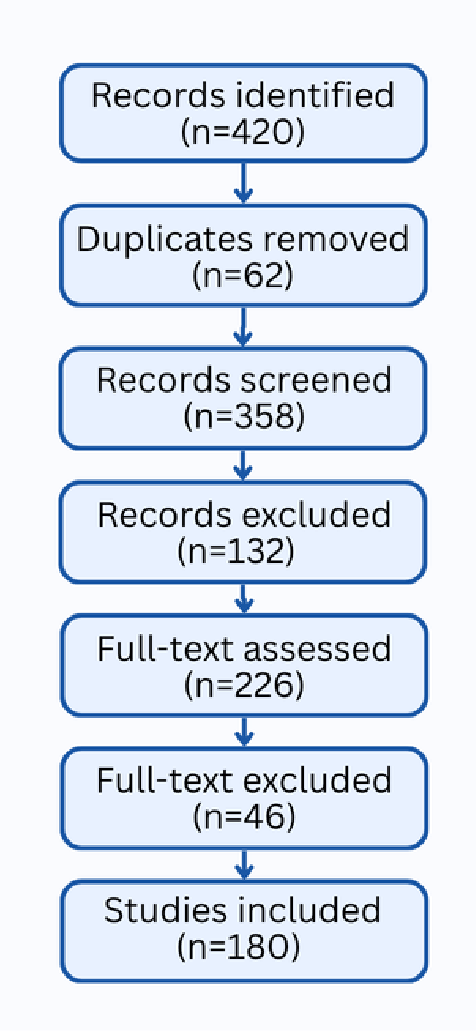

Figure 2. Overview of the study selection process used in the review, summarizing identification, screening, eligibility assessment, and final inclusion of articles.

the design of descriptors that captured clinically relevant lesion properties. Commonly used features included color histograms across RGB, HSV, and CIELab models; texture operators such as the Gray-Level Co-occurrence Matrix (GLCM) and Local Binary Patterns (LBP); and shape metrics aligned with the ABCD dermatological rule asymmetry, border irregularity, color variation, and diameter. In addition, multi-scale representations derived from wavelet and Gabor filters were employed to model fine-grained lesion characteristics [16].

Following feature extraction, classification was performed using conventional machine-learning algorithms such as Support Vector Machines (SVM), k-Nearest Neighbours (k-NN), Random Forests, Decision Trees, Logistic Regression, and Naïve Bayes. Among these, SVMs gained popularity in early melanoma studies because of their robustness on small datasets and relatively low overfitting tendency. However, their performance was highly sensitive to feature quality and segmentation accuracy, leading to brittle generalization across imaging devices, illumination conditions, and patient demographics [17].

Traditional computer-aided diagnosis (CAD) systems for skin disease detection are followed by a sequential workflow comprising dataset selection, preprocessing, segmentation, feature extraction, and classification. Each stage required careful manual design and optimization—preprocessing to remove artifacts such as hair and lighting variations, segmentation to delineate lesion boundaries, and handcrafted feature extraction based on color, texture, and shape descriptors. The overall pipeline is illustrated in Figure 3, summarizing the conventional stages that underpinned early systems and highlighting the need for more modular, automated deep-learning pipelines.

From a software-engineering perspective, these CAD pipelines resemble loosely coupled procedural programs rather than reusable systems. Each module—often written in isolation—required manual parameter tuning and lacked standardized interfaces for data exchange. Reproducing results across laboratories was therefore difficult, as preprocessing scripts, segmentation thresholds, and classifier parameters were rarely version-controlled or documented. Despite these shortcomings, traditional methods provided an architectural blueprint that later influenced the modular data pipelines seen in contemporary deep-learning frameworks.

Traditional CAD pipelines lacked the fundamental engineering practices required for scalable and reproducible systems. Each stage, i.e., preprocessing, segmentation, feature extraction, and classification, was implemented as a tightly coupled script with minimal modularity, making updates difficult and error-prone. Documentation was often incomplete; preprocessing steps were not version-controlled, and parameter settings were rarely archived, preventing exact replication across research groups. The absence of automated testing, standardized data interfaces, and pipeline orchestration meant that minor changes in one module frequently broke downstream components. These engineering limitations contributed to the shift toward deep learning, where end-to-end architectures reduce manual coupling and improve reproducibility through training pipelines.

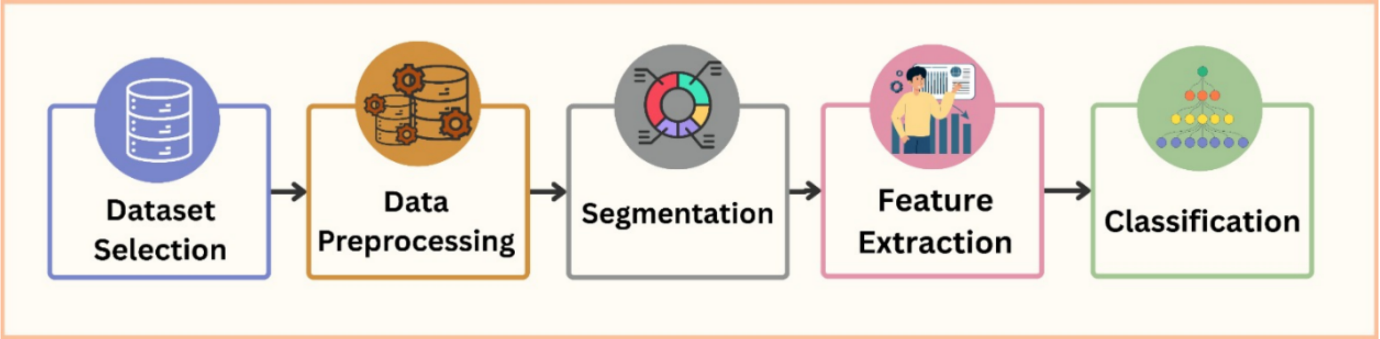

Figure 3. Traditional computer-aided diagnosis (CAD) workflow for skin disease detection, illustrating the sequential stages of dataset selection, data preprocessing, lesion segmentation, handcrafted feature extraction, and classical machine-learning-based classification.

Table 1. Traditional ML pipeline for skin disease classification and its limitations leading to deep learning.

Stage	**Techniques / Features**	**Algorithms / Methods**
Preprocessing	Hair removal (DullRazor), Gaussian filtering, color normalization	Image enhancement and artifact removal
Segmentation	Thresholding, region growing, watershed transformation, active contours	Delineating lesion boundaries
Feature Extraction (Handcrafted)	- Color histograms (RGB, HSV, CIELab) - Texture features: GLCM, LBP - Shape metrics (ABCD rule) - multi-scale features: Wavelet transforms, Gabor filters	Hand-designed descriptors to capture lesion characteristics
Classification	SVM, k-NN, Random Forests, Decision Trees, Logistic Regression, Naïve Bayes	Traditional ML classifiers for lesion categorization
Performance Examples	- SVM-based dermatofluoroscopy model: 95.8% sensitivity, 80.9% specificity on 214 lesions - HOG (~80%) outperformed Gabor (79%), SIFT (78%), and LBP (76%) on dermoscopy dataset	CAD accuracy limited by dataset size, handcrafted features, and segmentation errors
Limitations	- Dependency on handcrafted features - Poor generalization to diverse datasets - Segmentation errors propagated to classifiers - Dataset bias & spurious correlations- Absence of a unified, automated, and reproducible pipeline (key software-engineering limitation)	Motivated transition to deep learning

3.1 Performance & Limitations of Traditional Approaches

Before deep learning became mainstream, handcrafted feature–based CAD systems reported respectable but inconsistent results. For example, a dermatofluoroscopy study achieved 95.8 % sensitivity and 80.9 % specificity for melanoma detection using an SVM classifier on 214 pigmented lesions [18]. Another comparative study using a public dermoscopy dataset found that Histogram-of-Oriented-Gradients (HOG) features produced the highest single-classifier accuracy (˜80 %), outperforming Gabor (79 %), SIFT (78 %), and LBP (76 %) descriptors [19].

The comparative performance of classical handcrafted feature descriptors demonstrates the variability in their discriminative capacity for skin lesion classification. As shown in Figure 4, Histogram of Oriented Gradients (HOG) achieved the highest accuracy (˜80%), outperforming Gabor filters (79%), SIFT (78%), and Local Binary Patterns (LBP, 76%). These results underscore that while traditional methods could provide reasonable accuracy, their success was highly dependent on the quality of handcrafted features and the homogeneity of the dataset—factors that later motivated the transition toward deep learning–based approaches.

Yet these pipelines faced several engineering and methodological limitations. Most were trained on

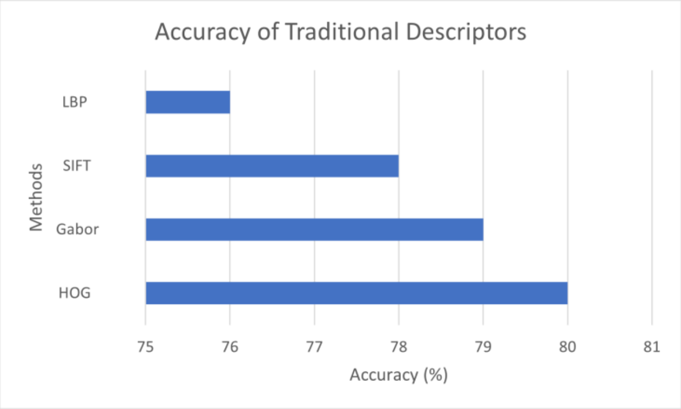

Figure 4. Comparative accuracy of traditional handcrafted feature descriptors used in early skin-disease classification systems, highlighting the performance variability across texture-, color-, and shape-based methods.

small, homogeneous datasets with minimal external validation, limiting scalability and cross-domain reliability. Handcrafted features were inherently shallow and often failed to capture the complex morphology of lesions, while segmentation errors propagated through the pipeline and degraded classification performance [20–22].

Moreover, recent analyses have revealed dataset bias and spurious correlations, where classical algorithms unintentionally learned background artifacts rather than true pathological features—raising questions about robustness and fairness in clinical deployment [4, 23]. While these methods achieved reasonable accuracy, their reliance on handcrafted features, segmentation quality, and limited datasets often restricted performance, paving the way for the shift toward deep learning are shown in Table 1.

In summary, classical CAD approaches laid the groundwork for data processing and feature standardization but lacked the engineering rigor—version control, modularity, automated testing, and deployment readiness—expected of contemporary AI systems. Their dependence on manual design and inability to self-adapt to unseen data underscored the need for end-to-end learnable and maintainable architectures, catalyzing the transition to deep learning and modern MLOps-oriented dermatology AI.

These limitations were not only algorithmic but also fundamentally software-engineering bottlenecks. Traditional pipelines required extensive manual parameter tuning for each stage, making them difficult to maintain. The lack of modularity meant that segmentation, feature extraction, and classification could not be updated or replaced independently, reducing extensibility and slowing development. Reproducing results was often impossible because preprocessing scripts, thresholds, and feature definitions were not standardized. Combined with poor cross-domain generalization arising from handcrafted features, these constraints created an unsustainable engineering model. Deep learning emerged as a response to these bottlenecks by offering end-to-end architectures that reduced manual dependencies, improved robustness, and supported more reproducible software workflows.

4 Machine and deep learning approaches

The transition from traditional computer-aided diagnosis (CAD) systems to deep learning–based architectures marked a decisive shift in how dermatological image analysis is conceived, implemented, and maintained. In software-engineering terms, this transformation represents a move from procedural, handcrafted pipelines toward end-to-end, data-driven systems with reusable components, defined interfaces, and automated learning of representations.

4.1 Evolution of Model Architectures

Early deep models adopted convolutional neural networks (CNNs) as the dominant architecture for learning hierarchical image features without explicit feature engineering. Networks such as AlexNet, VGGNet, and ResNet were retrained or fine-tuned on dermatology datasets to classify melanoma, basal-cell carcinoma, and other lesions [24]. These systems eliminated most manual preprocessing, achieving dermatologist-level accuracy on curated image sets. As datasets expanded—through initiatives such as the International Skin Imaging Collaboration (ISIC) and HAM10000—researchers began integrating transfer learning and data-augmentation pipelines to address class imbalance and overfitting. These practices introduced software-engineering discipline into experimental design, enabling reproducible training and evaluation across laboratories.

Subsequent work explored lightweight CNNs such as MobileNet, ShuffleNet, and SqueezeNet to facilitate deployment on mobile or embedded devices. These architectures emphasized computational efficiency, parameter sharing, and modularization, aligning naturally with SE goals of portability, scalability, and energy-aware design.

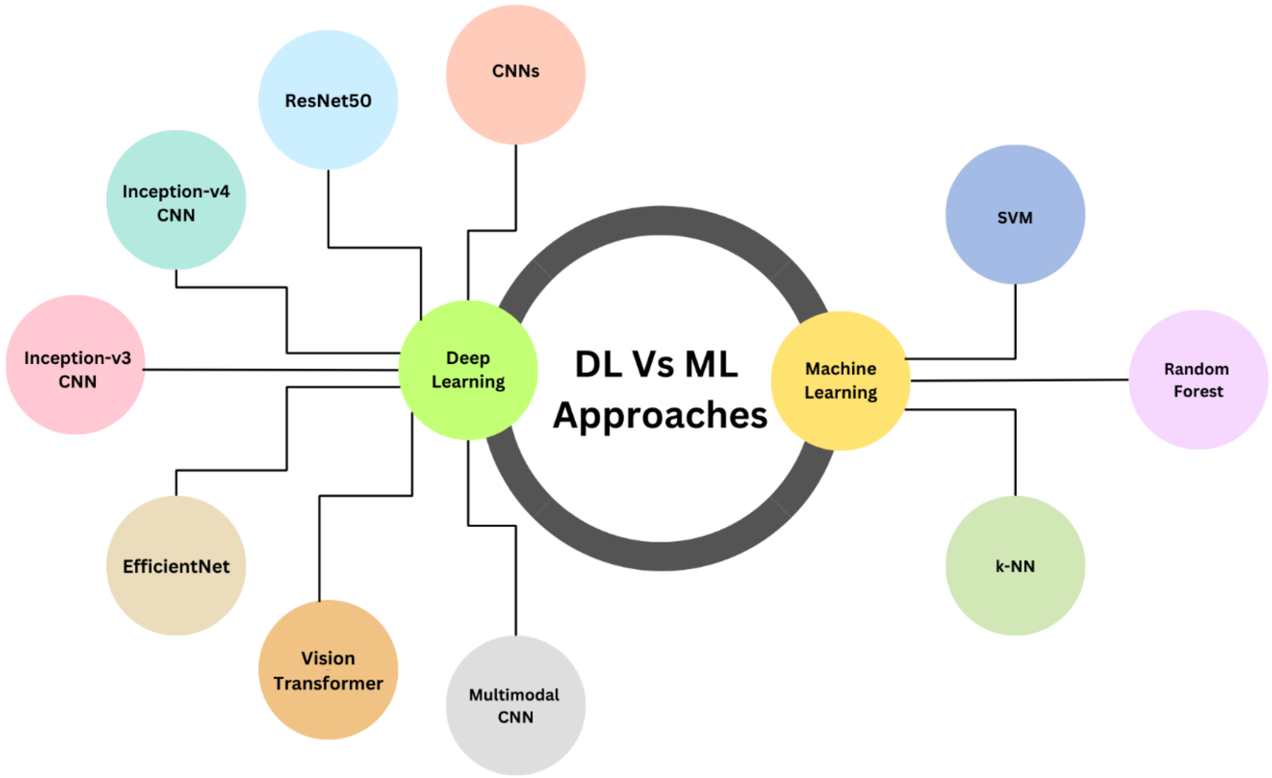

Figure 5. Representative deep-learning architectures for skin-disease classification, showing how model design has evolved toward more modular, maintainable, and deployment-ready software systems.

4.2 Advancements Beyond CNNs

Recent years (2021–2025) have witnessed a paradigm shift toward hybrid and transformer-based architectures that integrate image and metadata streams. Models such as EfficientNet, Vision Transformers (ViT), and Swin-Transformer introduced mechanisms for global context modeling and improved explainability through attention maps [24, 25]. Moreover, multimodal frameworks now combine dermoscopic images with structured clinical information—age, sex, lesion site, or skin tone—through dual-branch encoders that fuse visual and tabular data at the feature or decision level. From an SE perspective, these architectures behave as micro-services with defined APIs: each branch processes a specific modality and exposes standardized embeddings for downstream fusion. This modularity supports independent development, testing, and scaling—key properties of robust software systems. This means each module functions independently with a clear input and output, making the system easier to update, test, and maintain.

The Figure 5 contrasts early convolutional models (CNNs and EfficientNet) with transformer-based and multimodal architectures that integrate patient metadata through dual-branch pipelines. Each block reflects the increasing engineering maturity of dermatology AI—from static models to modular, API-driven components designed for reproducibility, scalability, and clinical deployment.

4.3 Learning Paradigms and Training Strategies

To overcome data scarcity and improve generalization, several studies introduced self-supervised learning and federated learning frameworks. Self-supervised pretraining exploits unlabeled images to learn domain-specific representations that transfer effectively to limited annotated datasets [27]. Federated learning extends this concept across institutions, enabling model training without centralizing patient data—thereby enhancing privacy, security, and compliance with regulations such as HIPAA and GDPR.

From an SE standpoint, these developments correspond to the DevOps to MLOps evolution: continuous integration and deployment of models, configuration management (e.g., Docker, Conda, or MLflow environments), and monitoring of model drift during clinical use. Tools such as TensorFlow Extended (TFX), Kubeflow, and MLflow have formalized reproducibility and traceability—dimensions now essential for medical-software certification.

4.4 Evaluation and Benchmarking Practices

Deep-learning studies typically evaluate model performance using metrics such as accuracy, precision, recall, F1-score, and area under the ROC curve (AUC). However, modern SE-aligned evaluations also include non-functional attributes:

- Latency and throughput for on-device inference
- Memory footprint and parameter count for deployment feasibility
- Calibration error (ECE or Brier score) for reliability
- Fairness and subgroup analysis across skin tones or imaging devices

Benchmark datasets such as ISIC, PH^2, Derm7pt, and BCN20000 have provided standardized test splits that promote comparability and regression testing across algorithm versions. Yet, inconsistencies remain in version control, seed initialization, and documentation—highlighting the ongoing need for reproducibility engineering within the dermatology AI community.

4.5 Interpretability and Human-AI Collaboration

Explainability has become integral to trustworthy AI. Approaches such as Grad-CAM, Layer-wise Relevance Propagation (LRP), and attention heatmaps visually link network activations to lesion regions, enabling clinicians to validate model reasoning. Embedding these interpretability modules as callable components in the inference pipeline enhances transparency, debuggability, and clinical acceptance.

Beyond interpretation, human-AI collaboration frameworks treat models as decision-support agents rather than replacements for dermatologists. Studies show that combining algorithmic predictions with expert review improves diagnostic accuracy and reduces inter-observer variability—aligning with SE principles of human-in-the-loop validation and safe system design.

4.6 Comparison of Machine Learning and Deep Learning Approaches

Between 2015 and 2025, dermatological image analysis witnessed a clear paradigm shift from classical machine learning (ML) to deep learning (DL)–based systems. Classical ML pipelines such as Support Vector Machines (SVMs), Random Forests (RF), and k-Nearest Neighbors (k-NN) relied on hand-engineered features—color histograms, texture descriptors, and shape metrics—to encode low-level lesion information. For instance, the work in [5] combined color and texture descriptors using SVMs and achieved 91.4 % accuracy on the PH^2 dataset, while [28] reported 85.7 % accuracy for handcrafted Gabor-based features with RF classifiers. Although these systems performed well on small curated datasets, they were constrained by their dependence on feature engineering and showed weak generalization across imaging devices, institutions, and populations.

The emergence of deep learning, particularly Convolutional Neural Networks (CNNs), introduced a data-driven alternative: hierarchical feature extraction learned directly from raw images. A landmark study employing Inception-v3 trained on over 129 450 dermoscopic images achieved dermatologist-level performance (AUC = 0.96) [7]. Subsequent architectures such as ResNet and EfficientNet further improved accuracy to 93–95 % on benchmark datasets including HAM10000 and ISIC 2019 [8]. These advances eliminated manual feature design, improved robustness to domain shifts, and enabled end-to-end optimization of the diagnostic pipeline.

In the later phase of this evolution, Transformer-based architectures began to redefine dermatology AI. Vision Transformers (ViTs), introduced around 2021, achieved 94.6 % accuracy on HAM10000 [29], offering improved global context modeling through self-attention. The first self-supervised and federated ViTs pushed performance to 96.1 % while preserving patient privacy [30]. These models embody a critical step toward software-engineered systems that integrate data privacy, model governance, and distributed learning within a unified MLOps framework.

From a software-engineering perspective, this evolution illustrates the shift from static pipelines to adaptive, maintainable systems characterized by:

- Automation: Feature extraction, training, and deployment orchestrated through reproducible pipelines.
- Scalability: Transfer-learning and lightweight models supporting mobile and edge deployment.
- Governance: Federated and privacy-aware learning ensuring regulatory compliance.
- Observability: Continuous monitoring of model drift, calibration, and fairness of metrics.

Table 2. Summary of ML and DL models for skin disease classification.

Study / Model	Dataset	Technique	Performance (Accuracy / AUC)	Key Findings / Limitations
SVM + Color & Texture [5]	PH²	Handcrafted features + SVM	91.4 % accuracy	Good on small datasets; poor cross-domain generalization.
RF + Gabor [28]	Dermoscopic images	Handcrafted features + RF	85.7 % accuracy	Feature engineering required; scalability issues.
k-NN + Custom Features [24]	Private dataset	Handcrafted + k-NN	~84 % accuracy	High variance; dataset dependency.
Inception-v3 CNN [7]	129 450 images	Deep CNN	AUC = 0.96	Dermatologist-level performance.
CNN Ensemble (ISIC) [25]	ISIC	CNN Ensemble	AUC = 0.91	Surpassed dermatologists in certain tasks.
ResNet [25]	HAM10000	Deep Residual Network	95 % accuracy	Strong benchmark; large data requirement.
ResNet-50 [8]	ISIC 2019	Deep CNN	95 % sens., 82 % spec.	Excellent melanoma detection; imbalanced data.
Inception-v4 [7]	Dermoscopic images	Deep CNN	AUC = 0.96	Reinforced dermatologist-level performance.
EfficientNet [26]	HAM10000	EfficientNet CNN	93.5 % accuracy	Better accuracy + efficiency balance.
Hybrid CNN + Metadata [10]	HAM10000 + metadata	Multimodal CNN	92.3 % accuracy	More robust via multimodal fusion.
Vision Transformer (ViT) [28, 29]	HAM10000	ViT	94.6 % accuracy	Global context modeling; high interpretability.
Self-Supervised ViT + Federated Learning [29, 31]	HAM10000 ISIC	ViT (SSL + FL)	96.1 % accuracy	Accuracy + privacy preservation; high compute cost.

A comparative summary of representative ML and DL models—including datasets, methods, and performance—is presented in Table 2, demonstrating how modern architectures outperform traditional systems not only in accuracy but also in engineering maturity and readiness for real-world clinical deployment.

In summary, the evolution of machine and deep-learning approaches for dermatology reflects a progressive software-engineering maturity curve:

- From monolithic scripts to modular, reusable architectures
- From static experiments to continuous, monitored pipelines (MLOps)
- From accuracy-only evaluation to multi-objective quality assessment encompassing performance, fairness, and reliability

These developments set the stage for the next section, which examines benchmark datasets and SE-grade evaluation protocols that support reproducibility and trustworthy deployment.

From a software-engineering standpoint, modern deep-learning approaches in dermatology increasingly depend on mature engineering workflows that ensure reliability and reproducibility. Continuous Integration and Continuous Deployment (CI/CD) pipelines support automated testing of preprocessing scripts, model updates, and evaluation of metrics, reducing the risk of silent failures between model versions. Containerization tools such as Docker and Singularity standardize training and inference environments, improving portability across institutions and hardware platforms. Model governance frameworks include MLflow tracking, version control for datasets and models, and audit logs that enable traceability and regulatory compliance, both essential for clinical AI. Moreover, SE quality attributes such as scalability, maintainability, and deployment readiness now directly affect model design choices, pushing researchers toward lightweight architectures, modular fusion pipelines, and robust MLOps ecosystems.

Table 3. Key public datasets for skin disease classification.

Dataset	Year	Images / Patients	Data Type	Notable Features / Use Case
PH^2 [5]	2013	200 dermoscopic images	Dermoscopy	Small, curated set; early benchmark for handcrafted-feature ML models.
ISIC Archive / Challenges [33]	2016–2020	1 million images	Dermoscopic + clinical	Standardized splits, public leaderboard, and reproducible evaluation scripts.
HAM10000 [34]	2018	10 015 images	Dermoscopy	Seven lesion classes; widely used for CNN benchmarking and model calibration.
Derm7pt [35]	2019	2 000 images	Dermoscopy + 7-point checklist	Supports explainable-AI and interpretability studies.
PAD-UFES-20 [36]	2020	1 379 patients / smartphone images	Clinical + mobile	Promotes domain-generalization and real-world evaluation.
SD-198 [37]	2021	6 584 images / 198 classes	Clinical photographs	Large-scale multi-class dataset; diverse lesion representation.
Multimodal HAM10000 + Metadata [10]	2022	10 015 + structured attributes	Dermoscopy + clinical data	Enables multimodal fusion and hybrid AI pipelines.

5 Datasets and Benchmarks for Skin Disease Classification

The creation of publicly accessible dermatology datasets and benchmark challenges between 2015 – 2025 has profoundly shaped how machine- and deep-learning systems are evaluated, reproduced, and compared. From a software-engineering standpoint, these datasets function as shared testbeds, enforcing standardized input–output specifications, metadata schemas, and versioned evaluation protocols. Together, they have transformed skin-disease classification from an isolated research activity into a reproducible software discipline.

Early studies relied on small, institution-specific datasets, such as the PH² dataset (200 dermoscopic images), which restricted model generalization and hindered external validation [5]. The subsequent emergence of large-scale benchmark repositories fundamentally changed the engineering of dermatology AI. The International Skin Imaging Collaboration (ISIC) Archive became the most widely used open resource, hosting millions of dermoscopic and clinical images. Through its annual ISIC Challenges (2016–2020), the archive introduced standardized train/validation/test splits, versioned evaluation scripts, and public leaderboards, providing the field with its first reproducible evaluation framework [33].

Another milestone was the HAM10000 dataset, released in 2018, comprising 10 015 dermoscopic images covering seven lesion categories. This dataset enabled training of deeper architectures such as ResNet and EfficientNet, achieving accuracies exceeding 93% on benchmark splits [34]. Similarly, the Derm7pt dataset offered 2 000 dermoscopic images annotated with 7-point melanoma checklist labels, encouraging explainable-AI research and clinical interpretability studies [35]. More recent datasets—PAD-UFES-20 and SD-198—introduced clinical and smartphone-acquired images, explicitly targeting domain generalization and cross-device robustness [35, 36].

Beyond pure image repositories, the field has begun integrating multimodal datasets that combine dermoscopic images with structured clinical metadata such as patient age, lesion site, and diagnostic history. Such datasets have driven the design of hybrid AI frameworks, enabling late- or cross-modal fusion of image and tabular streams [10]. These multimodal resources embody a software-system shift toward richer data models, supporting modular architectures and facilitating interoperability between image analysis engines and electronic health record (EHR) systems.

From an SE perspective, these datasets collectively serve three purposes:

1. Reproducibility — open access and versioned splits support verification of published results.

2. Bias Assessment — dataset documentation exposes over-representation of benign lesions and limited diversity across Fitzpatrick skin types.

3. Engineering Benchmarking — standardized protocols allow regression testing, cross-version validation, and monitoring of fairness and calibration in updated models.

A summary of the key public datasets—including their scale, modalities, and distinguishing attributes—is presented in Table 3, which highlights their pivotal role in reproducible software pipelines for dermatology AI.

In addition to the tabular summary, Figure 6 illustrates the chronological growth and increasing complexity of benchmark datasets used in dermatology AI.

6 Advanced Architectures in Skin Disease Classification (2022–2025)

Between 2022 and early 2025, dermatological AI has entered a phase characterized by architectural diversification and software-system maturity. The focus has shifted from single-modality convolutional neural networks (CNNs) toward efficient, interpretable, and multimodal deep learning frameworks, designed not only for accuracy but also for scalability, transparency, and deployability.

6.1 Lightweight and Efficient Transformers

One notable development is DermViT (2025)—a lightweight Vision Transformer (ViT) variant engineered for dermatology image analysis. DermViT introduces hierarchical attention mechanisms that suppress background noise and emphasize clinically relevant features, yielding 85.3% accuracy and a mean AUC (MAUC) of 96.3% on the ISIC-2018 and ISIC-2019 datasets. Its compact architecture demonstrates that transformer-based systems can achieve high diagnostic accuracy with minimal computational overhead, aligning with the principles of software efficiency and sustainability [38].

Similarly, SkinDistilViT (2023) represents a step forward in model distillation and resource optimization. Distilled from a large pre-trained teacher model, this network preserves 98.33% of its teacher's balanced multi-class accuracy on ISIC 2019 while reducing model size by 49.6% and achieving 69% faster GPU and 98% faster CPU inference [39]. These improvements exemplify engineering-driven design, where compression, latency reduction, and energy efficiency are core optimization goals alongside diagnostic accuracy.

6.2 Multimodal and Attention-Based Fusion Models

Recent research has expanded beyond vision-only architectures toward transformer-based multimodal fusion frameworks, which integrate image and metadata inputs for enhanced interpretability and robustness. One study proposed a one-stage attention-driven architecture capable of jointly reasoning over dermoscopic images and patient metadata [40]. This configuration is natively interpretable and outperforms traditional late-fusion pipelines by dynamically reweighting modalities according to contextual importance.

A major advance is the SkinM2Former model (2025)—a Tri-Modal Cross-Attention Transformer (TMCT) designed for multi-modal and multi-label classification. It jointly processes clinical images, dermoscopic images, and patient metadata, employing cross-attention layers to balance modality contributions. Different fusion strategies—early fusion, late fusion, and attention-based fusion—define

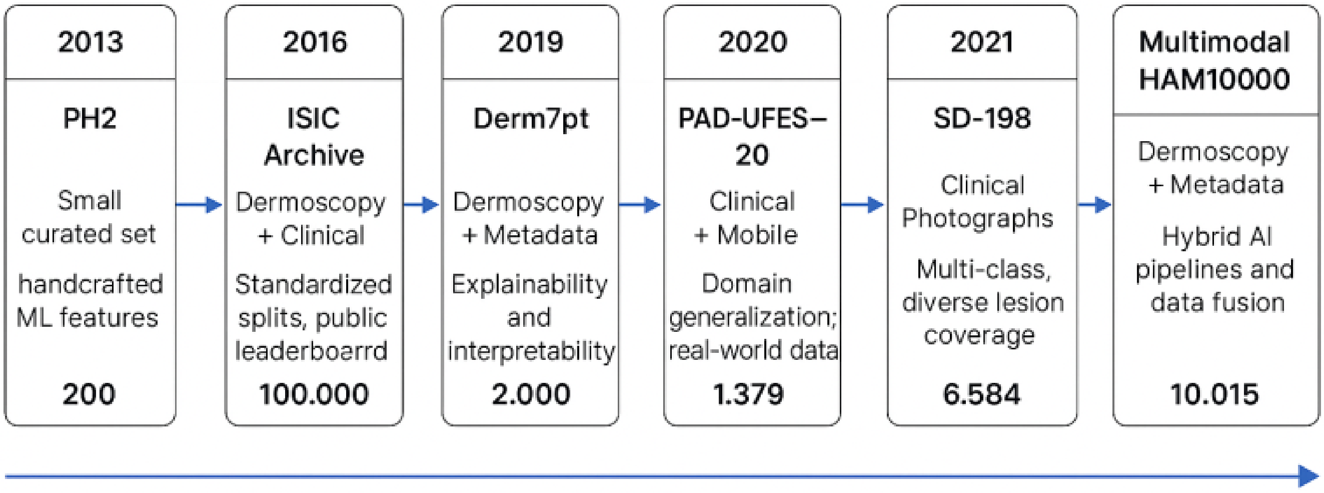

Figure 6. Growth of benchmark datasets for dermatology AI from 2013 to 2025, highlighting the expansion in dataset size, modality diversity, and engineering maturity—from small single-institution collections to large multimodal and mobile-acquired datasets.

Table 4. Recent transformer-based and multimodal approaches.

Model / Study	Technique / Key Innovation	Dataset(s)	Performance	Shortcomings/Limitations	Software-Engineering Attributes
DermViT	Hierarchical attention Vision Transformer (ViT); suppresses irrelevant background	ISIC-2018, ISIC-2019	85.3% accuracy, MAUC 96.3%	Performance lower than larger ViTs; not tested on multimodal data	Good maintainability due to modular attention blocks; scalable to larger datasets; reproducible architecture
SkinDistilViT	Lightweight ViT distilled from large teacher; efficient deployment	ISIC-2019	Retains 98.33% of teacher accuracy; 49.6% smaller; 69% faster (GPU), 98% faster (CPU)	Evaluated only on ISIC-2019; generalization not validated	Highly portable; excellent deployment readiness; ideal for CI/CD and edge/mobile environments
Cheslerean-Boghiu et al. [40]	Multimodal fusion (images + metadata) with attention; native interpretability	Private + benchmark datasets	Outperformed conventional fusion baselines	Dataset size limited; performance depends on metadata quality	Strong extensibility; requires structured data-governance; well-suited for modular fusion pipelines
SkinM2Former	Tri-Modal Cross-attention Transformer (TMCT) for multimodal, multilabel classification	Derm7pt	mAA 77.27%, Diagnostic accuracy 77.85%	High computational cost; tested only on Derm7pt	Supports modular multimodal design; scalable but requires high compute; maintainable cross-attention structure
PanDerm + ViT/Swin V2 Fusion	Dermatology-specific foundation model fused with ViTs	HAM10000, MSKCC	Fusion outperformed standalone ViTs	Foundation model training is resource-intensive; limited cross-dataset validation	Strong generalization potential; reusable components; foundation-model pipelines align with MLOps governance
BASNet + CCTM	Boundary-aware segmentation + CCTM classifier	PH^2, ISIC-2016/17/18, MED-NODE, HAM10000	IoU up to 0.98; Accuracy ~0.99	High complexity; limited interpretability; may hinder deployment	Robust but computationally heavy; requires careful optimization for deployment; maintainable due to separated segmentation/classifier modules
DinoV2 + Explainable AI	Transformer + GradCAM/SHAP interpretability	31-class dataset, HAM10000, Dermnet	Accuracy 96.48%; F1-score 0.9727	Explainability adds overhead; dataset-specific; limited clinical validation	High transparency; supports debugging and trust; integrates well with monitoring pipelines

how information is integrated across branches. On the Derm7pt dataset, SkinM2Former achieved a mean average accuracy (mAA) of 77.27% and a mean diagnostic accuracy (mDA) of 77.85%, outperforming prior state-of-the-art systems [41].

From an SE perspective, these designs illustrate how transformer architectures can be modularized into microservice-like components, enabling independent training, testing, and deployment of modality-specific branches. This modularity improves maintainability and supports continuous integration workflows common in clinical AI pipelines.

6.3 Ensemble and Hybrid Systems

Further innovation is seen in ensemble architectures that combine domain-specialized base models (PanDerm) with Vision Transformer (ViT). Studies from 2025 demonstrate that such ensembles consistently outperform single-model systems on datasets including HAM10000 and MSKCC, reinforcing the value of ensemble learning as a fault-tolerant software strategy [42].

Parallel work proposed a Boundary-Aware Segmentation Network (BASNet) combined with a Cross-Context Transformer Module (CCTM), achieving IoU scores up to 0.98 and classification accuracy approaching 0.99 across PH^2, ISIC (2016–2018), MED-NODE, and HAM10000 datasets [43]. These results underline the trend toward multi-task networks that unify segmentation and classification within a single, well-engineered system, improving both accuracy and computational reuse.

6.4 Explainable and Interpretable Transformers

The adoption of explainability frameworks such as Grad-CAM and SHAP has become central to enhancing clinical trust and system transparency. Grad-CAM visualizes activation heatmaps, revealing the lesion regions influencing each prediction, while SHAP attributes prediction weights to metadata features (e.g., patient age, lesion site). Integrated into transformer architectures like DinoV2, these methods enable both visual and semantic interpretability, essential for clinical auditing and human-AI collaboration.

Experiments conducted on a 31-class dataset, as well as on HAM10000 and DermNet, achieved 96.48% accuracy and an F1-score of 0.9727, confirming that transformer-based models can maintain both high accuracy and explainability in dermatology applications [44].

6.5 Summary and Emerging Trends

Taken together, these works signal a decisive shift toward multi-task learning, foundation-model distillation, multimodal fusion, and explainable transformer architectures. This convergence marks a new era in dermatology AI—where systems are not only accurate but also interpretable, maintainable, and resource-efficient. Collectively, they demonstrate how software-engineering principles—modularity, scalability, and transparency—are now embedded in the very architecture of medical AI. Recent transformer-based and multimodal approaches, summarizing datasets, techniques, and performance, are presented in Table 4.

Recent deep learning studies have advanced toward multimodal architectures that integrate multiple data streams for improved diagnostic performance. As illustrated in Figure 7, modern transformer-based systems combine clinical images, dermoscopic images, and patient metadata as inputs. These modalities are jointly processed through Vision Transformers (ViTs) equipped with cross-attention mechanisms, enabling the model to capture complementary information and contextual relationships across inputs. The resulting multi-label outputs enhance both diagnostic precision and interpretability, reflecting the growing convergence of computer vision and patient-centric data modeling in dermatology AI.

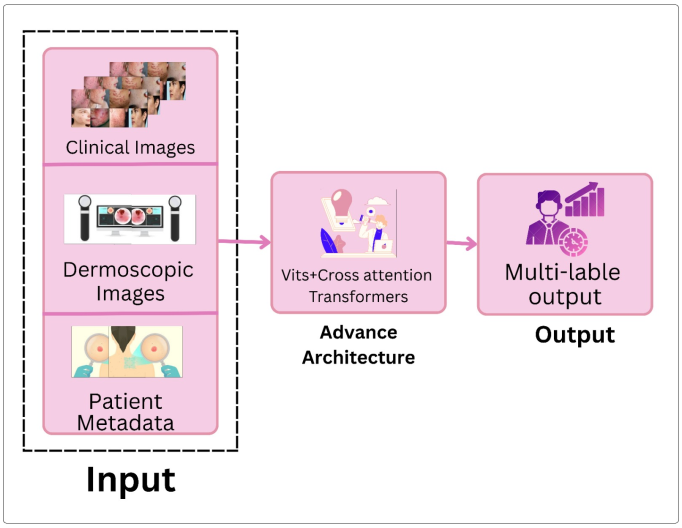

Figure 7. Multimodal transformer architecture for skin-disease classification, illustrating a modular design that integrates imaging and metadata streams to support scalable, maintainable clinical AI systems.

7 Challenges and Future Directions in Dermatology AI

Despite remarkable advances in artificial intelligence for dermatology, several engineering, ethical, and practical challenges continue to limit dependable clinical deployment. These challenges relate not only to data quality and algorithmic bias but also to software reliability, explainability, and governance—key concerns for trustworthy AI systems.

7.1 Class Imbalance and Data Bias

A persistent obstacle is class imbalance, where benign lesions dominate dermatology datasets while malignant cases, such as melanoma, remain comparatively rare. This imbalance can inflate overall accuracy while concealing poor sensitivity to critical conditions. Analyses of the ISIC challenges (2016–2020) demonstrate how uneven label distributions and evolving annotation standards make cross-year benchmarking difficult. Although methods such as data reweighting, focal loss, and balanced sampling partially mitigate these effects, they remain in software-level workarounds rather than systemic data solutions [45].

Bias across skin tones represents an equally pressing issue for fairness and equity. Many public datasets, including Fitzpatrick17k, contain limited examples of darker skin tones, resulting in models that underperform for under-represented groups. Recent work emphasizes inclusive data governance—moving beyond discrete skin-type labels toward continuous measures of hue, brightness, and texture diversity [45, 46]. Building balanced datasets with transparent subgroup reporting is therefore a software-engineering and ethical responsibility, not merely a data-preprocessing step.

7.2 Generalization and Domain Adaptation

Deep models often fail to generalize across clinical environments, imaging devices, and populations. A system trained in one hospital may underperform in another with different imaging protocols. Cross-dataset evaluations consistently show accuracy drops exceeding 10–15%, highlighting the fragility of current AI pipelines [49].

Promising approaches such as domain adaptation, active learning, and self-supervised pretraining have improved robustness, yet consistent generalization across sites remains elusive [47, 48]. Addressing this requires software-engineering principles such as version-controlled retraining, continuous monitoring, and automated calibration within MLOps frameworks.

7.3 Explainability and Clinical Trust

Explainability is central to responsible AI in dermatology. Clinicians must not only see what the model predicts but also understand why. While frameworks like CLEAR Derm advocate transparency, many visual explanation tools—such as Grad-CAM or LIME—fail to align with clinicians' cognitive reasoning. Recent studies show that when explanations mirror diagnostic logic (e.g., lesion border, texture, and asymmetry emphasis), they significantly improve human-AI collaboration [50, 51]. However, standardized metrics to evaluate the quality of explanations are still lacking. Future work should treat explainability as a software component—testable, auditable, and measurable—rather than an optional visualization feature.

7.4 Privacy-Preserving and Collaborative Learning

Cross-institutional collaboration is crucial for collecting large and diverse datasets while preserving patient confidentiality. Federated learning and hybrid model-sharing frameworks enable distributed training without transferring sensitive data. Such systems have achieved competitive results on ISIC and HAM10000, demonstrating the promise of privacy-preserving collaboration [32, 51–53].

Nonetheless, challenges like communication efficiency, model drift, and governance of federated nodes persist. These require coordinated efforts in software architecture design, ensuring that distributed systems are both secure and auditable.

7.5 Future Directions

Looking ahead, the transition from experimental success to trustworthy clinical deployment will depend on advances that integrate ethical awareness with software-engineering discipline. Future dermatology AI systems must be developed on intentionally diverse datasets, built through transparent data governance that documents subgroup composition and reporting practices. Fairness should be embedded directly into model design using reweighting, adversarial debiasing, and calibrated decision thresholds, rather than treated as a post hoc correction. Equally important is robust evaluation across hospitals, imaging devices, and populations, with external validation and domain-shift testing serving as default requirements before deployment. Beyond performance metrics like AUC, reader studies and prospective clinical trials should quantify how AI affects clinician accuracy, confidence, and workflow integration. The field also requires standardized reporting and monitoring frameworks, such as CLEAR Derm, to ensure model calibration, drift detection, and fail-safe recovery are part of every operational pipeline. Collectively, these steps will transform dermatology AI from isolated algorithmic achievements into reliable, fair, and reproducible clinical software systems, capable of complementing human expertise in real-world healthcare settings.

The challenges facing dermatology AI can be conceptualized across three interconnected levels—data, model, and deployment as shown in Figure 8. At the foundation lie data challenges, including class imbalance, underrepresentation of darker skin tones, and limited cross-domain generalization. These affect all subsequent stages of model training and evaluation. The middle layer consists of model challenges, where explainability and robustness remain critical for building clinician trust and ensuring resilience under varied imaging conditions. At the apex are deployment challenges, encompassing privacy preservation, clinical validation, and regulatory compliance. Addressing each tier requires a coordinated software-engineering approach—standardized data governance, reproducible model pipelines, and secure, auditable deployment frameworks—that together enable trustworthy, equitable, and clinically viable AI systems.

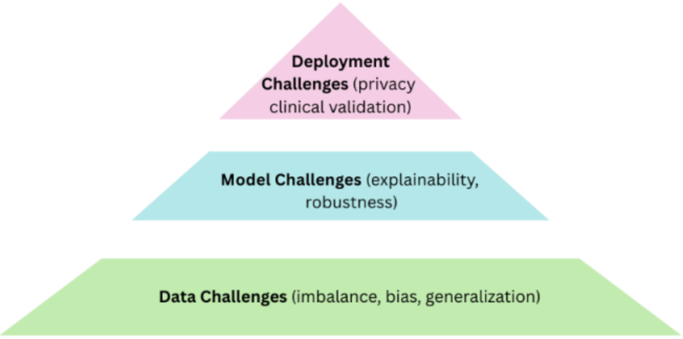
Deployment Challenges (privacy clinical validation)
Model Challenges (explainability, robustness)
Data Challenges (imbalance, bias, generalization)

Figure 8. Hierarchical challenges in dermatology AI. The pyramid illustrates how data, model, and deployment challenges collectively influence reliability—spanning imbalance and bias at the data level, explainability and robustness at the model level, and privacy and clinical validation at deployment.

7.6 End-to-End SE Workflow

An end-to-end dermatology AI workflow can be understood as a software system composed of modular, testable, and continuously monitored components. The process begins with dataset acquisition and versioning, where raw images and metadata are tracked using tools such as DVC or Git-LFS to ensure traceability. Preprocessing pipelines—including resizing, normalization, and augmentation—are implemented as reproducible modules supported by automated validation tests. Model development and training are orchestrated through containerized environments (e.g., Docker), ensuring consistent execution across machines and institutions. During deployment, trained models are exposed through optimized REST APIs or containerized microservices and integrated into a CI/CD pipeline for automated validation and controlled rollout. After deployment, monitoring systems track performance drift, latency, reliability, and clinical feedback, enabling continuous updates while preserving regulatory and safety requirements. This workflow demonstrates how dermatology AI aligns with a full software-engineering lifecycle rather than an isolated machine-learning model.

8 Conclusion

The rapid integration of machine learning and deep learning has significantly advanced skin disease classification, with convolutional neural networks (CNNs), vision transformers (ViTs), and hybrid models consistently outperforming traditional approaches such as SVMs, RF, and k-NN. These advances demonstrate clear improvements in diagnostic accuracy, robustness, and scalability, highlighting a paradigm shift toward data-driven dermatology. However, critical challenges remain unresolved, including class imbalance, bias across skin tones, limited domain generalization, and the lack of reliable explainability frameworks. Without addressing these issues, widespread clinical adoption may risk reinforcing inequities and reducing trust among practitioners. Looking ahead, promising directions such as federated learning, fairness-aware AI, robust domain adaptation, and explainable frameworks offer pathways to trustworthy and equitable deployment. Equally important is the shift from retrospective benchmarks to prospective validation through clinician AI collaboration and rigorous reporting standards. By combining technical innovation with fairness, transparency, and clinical validation, skin disease classification systems can evolve from research prototypes into reliable, real-world diagnostic support tools that improve healthcare outcomes across diverse populations. Overall, the findings of this review demonstrate that the clinical usefulness of advanced dermatology AI models depends fundamentally on strong software-engineering practices, including reproducibility, maintainability, and robust deployment workflows.

Data Availability Statement

Data will be made available on request.

Funding

This work was supported without any funding.

Conflicts of Interest

The authors declare no conflicts of interest.

AI Use Statement

The authors declare that no generative AI was used in the preparation of this manuscript.

Ethical Approval and Consent to Participate

Not applicable. This is a review article based on publicly available literature and datasets; no human participants, primary data collection, or clinical interventions were involved.

References

[1] He, M., & Zhang, X. (2024, November). A Review of Research Advances in Image Segmentation of Skin Lesions. In *International Artificial Intelligence Conference* (pp. 265-279). Singapore: Springer Nature Singapore. [CrossRef]

[2] Daneshjou, R., Barata, C., Betz-Stablein, B., Celebi, M. E., Codella, N., Combalia, M., ... & Rotemberg, V. (2022). 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 dermatology, 158*(1), 90-96. [CrossRef]

[3] Karimkhani, C., Boyers, L. N., Dellavalle, R. P., & Weinstock, M. A. (2015). It's time for 'keratinocyte carcinoma' to replace the term 'nonmelanoma skin cancer'. *Journal of the American Academy of Dermatology, 72*(1), 186–187. [CrossRef]

[4] Groh, M., Harris, C., Soenksen, L., Lau, F., Han, R., Kim, A., ... & Badri, O. (2021). Evaluating deep neural networks trained on clinical images in dermatology with the fitzpatrick 17k dataset. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition* (pp. 1820-1828).

[5] Barata, C., Celebi, M. E., & Marques, J. S. (2015). Improving dermoscopy image classification using color constancy. *IEEE Journal of Biomedical and Health Informatics, 19*(3), 1146–1152. [CrossRef]

[6] Cassidy, B., Kendrick, C., Brodzicki, A., Jaworek-Korjakowska, J., & Yap, M. H. (2022). Analysis of the ISIC image datasets: Usage, benchmarks and recommendations. *Medical image analysis, 75*, 102305. [CrossRef]

[7] Esteva, A., Kuprel, B., Novoa, R. A., Ko, J., Swetter, S. M., Blau, H. M., & Thrun, S. (2017). Dermatologist-level classification of skin cancer with deep neural networks. *Nature, 542*(7639), 115–118. [CrossRef]

[8] Brinker, T. J., Hekler, A., Enk, A. H., Berking, C., Haferkamp, S., Hauschild, A., ... & Schadendorf, D. (2019). Deep learning outperformed 136 of 157 dermatologists in a head-to-head dermoscopic melanoma image classification task. *European Journal of Cancer, 113*, 47–54. [CrossRef]

[9] Codella, N., Rotemberg, V., Tschandl, P., Celebi, M. E., Dusza, S., Gutman, D., ... & Halpern, A. (2019). Skin lesion analysis toward melanoma detection 2018: A challenge hosted by the international skin imaging collaboration (isic). *arXiv preprint arXiv:1902.03368*.

[10] Gessert, N., Nielsen, M., Shaikh, M., Werner, R., & Schlaefer, A. (2020). Skin lesion classification using ensembles of multi-resolution EfficientNets with meta data. *MethodsX, 7*, 100864. [CrossRef]

[11] Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., ... & Polosukhin, I. (2017). Attention is all you need. In *Advances in neural information processing systems 30* (*NeurIPS 2017*).

[12] Hosny, K. M., Kassem, M. A., & Foaud, M. M. (2018). Skin cancer classification using deep learning and transfer learning. In *2018 9th Cairo International Biomedical Engineering Conference* (*CIBEC*) (pp. 90–93). IEEE. [CrossRef]

[13] Daneshjou, R., Barata, C., Betz-Stablein, B., Celebi, M. E., Codella, N., Combalia, M., ... & Rotemberg, V. (2022). 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 Dermatology, 158*(1), 90–96. [CrossRef]

[14] Alipour, N., Burke, T., & Courtney, J. (2024). Skin type diversity in skin lesion datasets: a review. *Current Dermatology Reports, 13*(3), 198-210. [CrossRef]

[15] Wu, L., & Tao, T. (2025). MARTE-based modeling and analysis for real-time neuromorphic computing in embedded systems. *ICCK Journal of Software Engineering, 1*(1), 9-16. [CrossRef]

[16] Debelee, T. G. (2023). Skin lesion classification and detection using machine learning techniques: A systematic review. *Diagnostics, 13*(19), 3147. [CrossRef]

[17] Celebi, M. E., Wen, Q., Hwang, S., Iyatomi, H., & Schaefer, G. (2013). Lesion border detection in dermoscopy images using ensembles of thresholding methods. *Skin Research and Technology, 19*(1), e252-258. [CrossRef]

[18] Szyc, Ł., Hillen, U., Scharlach, C., Kauer, F., & Garbe, C. (2019). Diagnostic performance of a support vector machine for dermatofluoroscopic melanoma recognition: The results of the retrospective clinical study on 214 pigmented skin lesions. *Diagnostics, 9*(3), 103. [CrossRef]

[19] Zahid, M., Rziza, M., & Alaoui, R. (2025). Skin lesion classification using hybrid feature extraction based on classical and deep learning methods. *BioMedInformatics, 5*(3), 41. [CrossRef]

[20] Jeong, H. K., Park, C., Henao, R., & Kheterpal, M. (2023). Deep learning in dermatology: A systematic review of current approaches, outcomes, and limitations. *JID Innovations, 3*(1), 100150. [CrossRef]

[21] Zhang, J., Zhong, F., He, K., Ji, M., Li, S., & Li, C. (2023). Recent advancements and perspectives in the diagnosis of skin diseases using machine learning and deep learning: A review. *Diagnostics, 13*(23), 3506. [CrossRef]

[22] Fogelberg, K., Chamarthi, S., Maron, R. C., Niebling, J., & Brinker, T. J. (2023). Domain shifts in dermoscopic skin cancer datasets: Evaluation of essential limitations for clinical translation. *New Biotechnology, 76*, 106-117. [CrossRef]

[23] Bissoto, A., Fornaciali, M., Valle, E., & Avila, S. (2019). (De)constructing bias on skin lesion datasets. In *2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops* (*CVPRW*) (pp. 2766–2774). IEEE. [CrossRef]

[24] Puri, P., Combalia, M., Rotemberg, V., Carrera, C., Puig, S., Malvehy, J., ... & Tschandl, P. (2022). Deep learning for dermatologists: Part II. Current applications. *Journal of the American Academy of Dermatology, 87*(6), 1352–1360. [CrossRef]

[25] Jaworek-Korjakowska, J., & Kleczek, P. (2018). Eskin: study on the smartphone application for early detection of malignant melanoma. *Wireless Communications and Mobile Computing, 2018*(1), 5767360. [CrossRef]

[26] Liu, Y., Jain, A., Eng, C., Way, D. H., Lee, K., Bui, P., ... & Coz, D. (2020). A deep learning system for differential diagnosis of skin diseases. *Nature Medicine, 26*(6), 900–908. [CrossRef]

[27] Jin, C., Guo, Z., Lin, Y., Luo, L., & Chen, H.

(2023). Label-efficient deep learning in medical image analysis: Challenges and future directions. *arXiv preprint arXiv:2303.12484*.

[28] Kawahara, J., BenTaieb, A., & Hamarneh, G. (2016). Deep features to classify skin lesions. In *2016 IEEE 13th International Symposium on Biomedical Imaging (ISBI)* (pp. 1397–1400). IEEE. [CrossRef]

[29] Srinivas, A., Lin, T. Y., Parmar, N., Shlens, J., Abbeel, P., & Vaswani, A. (2021, June). Bottleneck Transformers for Visual Recognition. In *2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)* (pp. 16514-16524). IEEE. [CrossRef]

[30] Lin, H., Xu, C., & Qin, J. (2025). Taming Vision-Language Models for Medical Image Analysis: A Comprehensive Review. *arXiv preprint arXiv:2506.18378*.

[31] Du, S., Bayasi, N., Hamarneh, G., & Garbi, R. (2023, October). Avit: Adapting vision transformers for small skin lesion segmentation datasets. In *International Conference on Medical Image Computing and Computer-Assisted Intervention* (pp. 25-36). Cham: Springer Nature Switzerland. [CrossRef]

[32] Zhang, F., Yuan, K., Li, X., Gao, Y., Liu, Y., Wang, Z., ... & Zhang, D. (2025). Federated cross-incremental self-supervised learning for medical image segmentation. *IEEE Transactions on Neural Networks and Learning Systems, 36*(7), 13498–13511. [CrossRef]

[33] Mendonca, T., Ferreira, P. M., Marques, J. S., Marcal, A. R. S., & Rozeira, J. (2013). PH2 - a dermoscopic image database for research and benchmarking. In *2013 35th Annual International Conference of the IEEE Engineering in Medicine and Biology Society (EMBC)* (pp. 5437–5440). IEEE. [CrossRef]

[34] Tschandl, P., Rosendahl, C., & Kittler, H. (2018). The HAM10000 dataset, a large collection of multi-source dermatoscopic images of common pigmented skin lesions. *Scientific data, 5*(1), 1-9. [CrossRef]

[35] Combalia, M., Codella, N. C., Rotemberg, V., Helba, B., Vilaplana, V., Reiter, O., ... & Malvehy, J. (2019). Bcn20000: Dermoscopic lesions in the wild. *arXiv preprint arXiv:1908.02288*.

[36] Dong, C., Dai, D., Zhang, Y., Zhang, C., Li, Z., & Xu, S. (2023). Learning from dermoscopic images in association with clinical metadata for skin lesion segmentation and classification. *Computers in Biology and Medicine, 152*, 106321. [CrossRef]

[37] Pacheco, A. G. C., Lima, G. R., Salomao, A. S., Krohling, B., Biral, I. P., de Angelo, G. G., ... & Krohling, R. A. (2020). PAD-UFES-20: A skin lesion dataset composed of patient data and clinical images collected from smartphones. *Data in Brief, 32*, 106221. [CrossRef]

[38] Zhang, X., Song, C., Li, S., Wang, Y., Liu, J., Liu, L., ... & He, X. (2025). DermViT: Diagnosis-guided vision transformer for robust and efficient skin lesion classification. *Bioengineering, 12*(4), 421. [CrossRef]

[39] Lungu-Stan, V. C., Cercel, D. C., & Pop, F. (2023, September). Skindistilvit: Lightweight vision transformer for skin lesion classification. In *International Conference on Artificial Neural Networks* (pp. 268-280). Cham: Springer Nature Switzerland. [CrossRef]

[40] Cheslerean-Boghiu, T., Fleischmann, M. E., Willem, T., & Lasser, T. (2023). Transformer-based interpretable multi-modal data fusion for skin lesion classification. *arXiv preprint arXiv:2304.14505*.

[41] Zhang, Y., Xie, Y., Wang, H., Avery, J. C., Hull, M. L., & Carneiro, G. (2025, February). A Novel Perspective for Multi-modal Multi-label Skin Lesion Classification. In *2025 IEEE/CVF Winter Conference on Applications of Computer Vision (WACV)* (pp. 3549-3558). IEEE. [CrossRef]

[42] Mahbod, A., Ecker, R., & Woitek, R. (2025). Fusion of Foundation and Vision Transformer Model Features for Dermatoscopic Image Classification. *arXiv preprint arXiv:2505.16338*.

[43] Amin, J., Azhar, M., Arshad, H., Zafar, A., & Kim, S. H. (2025). Skin-lesion segmentation using boundary-aware segmentation network and classification based on a mixture of convolutional and transformer neural networks. *Frontiers in Medicine, 12*, 1524146. [CrossRef]

[44] Mohan, J., Sivasubramanian, A., & Ravi, V. (2025). Enhancing skin disease classification leveraging transformer-based deep learning architectures and explainable ai. *Computers in Biology and Medicine, 190*, 110007. [CrossRef]

[45] Cassidy, B., Kendrick, C., Brodzicki, A., Jaworek-Korjakowska, J., & Yap, M. H. (2022). Analysis of the ISIC image datasets: Usage, benchmarks and recommendations. *Medical Image Analysis, 75*, 102305. [CrossRef]

[46] Groh, M., Harris, C., Soenksen, L., Lau, F., Han, R., Kim, A., ... & Badri, O. (2021, June). Evaluating Deep Neural Networks Trained on Clinical Images in Dermatology with the Fitzpatrick 17k Dataset. In *2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)* (pp. 1820-1828). IEEE. [CrossRef]

[47] Alipour, N., Burke, T., & Courtney, J. (2024). Skin type diversity in skin lesion datasets: A review. *Current Dermatology Reports, 13*(3), 198–210. [CrossRef]

[48] Montoya, L. N., Roberts, J. S., & Hidalgo, B. S. (2025, March). Towards fairness in AI for melanoma detection: Systemic review and recommendations. In *Future of Information and Communication Conference* (pp. 320-341). Cham: Springer Nature Switzerland. [CrossRef]

[49] Fogelberg, K., Chamarthi, S., Maron, R. C., Niebling, J., & Brinker, T. J. (2023). Domain shifts in dermoscopic skin cancer datasets: Evaluation of essential limitations for clinical translation. *New*

Biotechnology, 76, 106–117. [CrossRef]

[50] Ye, J. (2025, July). Enhancing skin lesion classification generalization with active domain adaptation. In *2025 47th Annual International Conference of the IEEE Engineering in Medicine and Biology Society (EMBC)* (pp. 1-7). IEEE. [CrossRef]

[51] Zakhem, G. A., Fakhoury, J. W., Motosko, C. C., & Ho, R. S. (2021). Characterizing the role of dermatologists in developing artificial intelligence for assessment of skin cancer. *Journal of the American Academy of Dermatology, 85*(6), 1544-1556. [CrossRef]

[52] Chanda, T., Haggenmueller, S., Bucher, T. C., Holland-Letz, T., Kittler, H., Tschandl, P., ... & Brinker, T. J. (2025). Dermatologist-like explainable AI enhances melanoma diagnosis accuracy: eye-tracking study. *Nature Communications, 16*(1), 4739. [CrossRef]

[53] Hendrix, R., Proietto Salanitri, F., Spampinato, C., Palazzo, S., & Bagci, U. (2024, December). Evidential Federated Learning for Skin Lesion Image Classification. In *International Conference on Pattern Recognition* (pp. 354-365). Cham: Springer Nature Switzerland. [CrossRef]

Moomna Nazir holds a bachelor's degree in computer science from COMSATS University Islamabad, Pakistan. Her academic interests focus on machine learning, IoT systems, and software design. She is passionate about applying practical security methods to real-world computing environments and continues to explore research opportunities in the field of emerging technologies. (Email: moomnanazir@gmail.com)

Azka Ahsan is affiliated with the Department of Computer Science at COMSATS University Islamabad, Sahiwal Campus. Her research interests span artificial intelligence, data analytics, and automated software systems. She is particularly focused on developing intelligent solutions that enhance system reliability and automation. (Email: azkaahsan915@gmail.com)

Rabia Khadim is a researcher at the Department of Computer Science, COMSATS University Islamabad, Sahiwal Campus. Her areas of interest include deep learning, computer vision, and human–computer interaction. She is committed to advancing AI-driven applications for healthcare and public-sector systems. (Email: rabiakhadim559@gmail.com)

Shakeel Abbas is associated with the Department of Computer Science, Minhaj University Lahore. His research areas include distributed systems, cloud computing, and secure software architectures. He is actively engaged in developing scalable and secure computational models for modern enterprise environments. (Email: shakeelabbas33@gmail.com)

Aown Muhammad is with the Department of Computer Science, University of Engineering and Technology (UET), Lahore. His research focuses on artificial intelligence, embedded systems, and intelligent automation. He has a strong interest in building efficient and resource-aware systems for industrial and academic applications. (Email: mail@aown.me)

Zain Sohail is affiliated with the Department of Computer Science at COMSATS University Islamabad, Sahiwal Campus. His research interests include machine learning, mobile computing, and intelligent systems. He is dedicated to leveraging data-driven technologies to solve societal and engineering challenges. (Email: zainsohail913@gmail.com)