RAMReS Sciences des Structures et de la Matière Vol. 9, N°2 (2025) 40 - 53
http://publication.lecames.org/index.php/mat ISSN: 2630-1180

Literature review on classification model for human skin disease

T. Ferdinand Guinko[1,*], B. Tanguy Kaboré[1,2]

[1]*Information Science and Knowledge Engineering, Laboratory of Mathematics and Computer Science, Joseph KI-ZERBO University, Ouagadougou, Burkina Faso*
[2] *Mathematical and Applied Computer Science Laboratory, Norbert ZONGO University, Koudougou, Burkina Faso*

Received: 30 May 2025 / Received in revised form: 19 October 2025 / Accepted: 12 December 2025

***Abstract*:**

The rising global burden of skin diseases requires precise and efficient diagnostic tools to improve patient care and treatment outcomes. Recent advances in artificial intelligence, particularly deep learning, have significantly contributed to the automation of the classification of skin diseases. This review presents a detailed analysis of various classification models, including ResNet, VGG, DenseNet, EfficientNet, and Transformers, evaluating their performance based on accuracy, recall, and the F1-score. Although traditional CNN architectures remain effective, the rapid advancement of AI models underscores their growing limitations and the risk of technological obsolescence. Novel approaches, such as Transformers and hybrid deep learning frameworks, show promise in achieving superior diagnostic performance while optimizing computational efficiency. By critically addressing the strengths and limitations of existing classification models, this review offers valuable insights for researchers and dermatology professionals, facilitating the adoption of state-of-the-art AI-driven diagnostic solutions.

***Keywords*:** Skin diseases; Classification model; Artificial intelligence; Deep learning; Convolutional neural network.

1. Introduction

Skin diseases represent a significant public health concern, affecting an estimated 1.8 billion people worldwide, as highlighted during the first World Health Organization (WHO) Global Meeting on Cutaneous Neglected Tropical Diseases (cNTDs) in March 2023 [1]. In tropical and resource-limited regions, bacterial,

* Corresponding author:
Email address: *fguinko@ujkz.bf (T.F. Guinko)*
https://doi.org/10.70974/mat0922540

viral, fungal, and parasitic skin infections remain among the most prevalent health conditions. Additionally, cNTDs account for approximately 10% of all skin diseases, further emphasizing the need for robust diagnostic solutions. Given the growing prevalence of skin conditions, early and accurate diagnosis is crucial to ensure effective treatment and management. In response to this challenge, artificial intelligence (AI), particularly deep learning, has emerged as a transformative tool for automating dermatological diagnosis. Over the past decade, convolutional neural networks (CNNs) have been widely applied to the classification of skin diseases, demonstrating considerable success in clinical and mobile health applications. However, with the rapid evolution of AI technologies, newer architectures such as EfficientNet, Transformers, and hybrid deep learning models are being developed to optimize both accuracy and computational efficiency.

Despite the progress in AI-driven skin disease classification, a major challenge persists due to the obsolescence of traditional classification models resulting from rapid advancements in deep learning. This review addresses the evolving landscape of AI-based skin disease classification, critically analyzing the strengths and limitations of existing models. By doing so, it seeks to highlight the need for continuous model updates and the adoption of more robust AI solutions to ensure accurate, scalable, and interpretable dermatological diagnosis.

This review explores recent advances in AI-driven skin disease classification, analyzing the most widely used models, their performance on different datasets, and their potential to revolutionize dermatological diagnostics. By identifying key trends and limitations, this study aims to provide a comprehensive resource for researchers and healthcare professionals, facilitating the transition to more advanced and efficient AI-based diagnostic systems.

2. Methods

As part of our literature review, we used a tool called *Right Review*, which guided us in selecting a quantitative systematic review as the most appropriate approach for our research objectives. This method, known for its rigor and reproducibility, enables a structured synthesis of empirical evidence and facilitates quantitative comparisons between studies. Consequently, we adopted the Preferred Reporting Items for Systematic Reviews and Meta-Analyses (PRISMA) framework, widely recognized as the gold standard for systematic literature reviews in medical and AI research.

2.1. Keywords

To ensure comprehensive coverage of relevant literature, we defined a set of precise and representative keywords aligned with our research objectives. These included: Skin diseases, Black skin diseases, Segmentation, Classification, Artificial intelligence, Machine learning, Deep learning, and Convolutional neural network. These keywords were selected based on their frequency and relevance in dermatological AI studies published between 2023 and 2025.

2.2. Search equations

Based on these keywords, we formulated a general search equation that combines terms for skin diseases, segmentation or classification tasks, and artificial intelligence methods, as follows: “Skin diseases” OR “Black skin diseases” AND (Segmentation OR Classification) AND (“Artificial intelligence” OR “Machine learning” OR “Deep learning” OR “Convolutional neural network”).

This query was adapted to the syntax requirements of each database to capture a broad range of studies focusing on automatic skin disease detection, segmentation, and classification using AI techniques.

2.3. Article selection process

Our systematic search covered four major academic databases: Google Scholar,

PubMed, Scopus, and Semantic Scholar. In total, 153 studies were initially identified (Google Scholar: 21; PubMed: 3; Scopus: 55; Semantic Scholar: 74). After removing 10 duplicates, 143 records were screened. During the screening phase, 14 studies were excluded based on title and abstract relevance, while 107 could not be retrieved in full text. Of the remaining 22 studies, 7 were excluded for irrelevance to our inclusion criteria. Ultimately, 15 studies were retained for detailed analysis. Each selected study was examined using a standardized extraction grid that included:

- The architecture type (e.g., CNN, Transformer, hybrid);
- dataset characteristics (source, size, Fitzpatrick type distribution);
- evaluation metrics (accuracy, precision, recall, F1-score);
- computational efficiency (inference time, model size, FLOPs).

To ensure comparability, we prioritized studies reporting standardized metrics (accuracy, precision, recall, F1-score) on well-established benchmark datasets (e.g., ISIC2016-2020, HAM10000). When multiple metrics were available, we selected those obtained on validation or test sets to avoid overfitting bias. For models evaluated on the same dataset, direct performance comparisons were made; for different datasets, we contextualized results based on dataset complexity and class distribution. This structured evaluation forms the basis of our comparative analysis and supports the claims discussed in the Discussion and Conclusion sections. The selection and filtering process is illustrated in Figure 1, which presents the PRISMA flow diagram corresponding to this systematic review on AI-based skin disease classification.

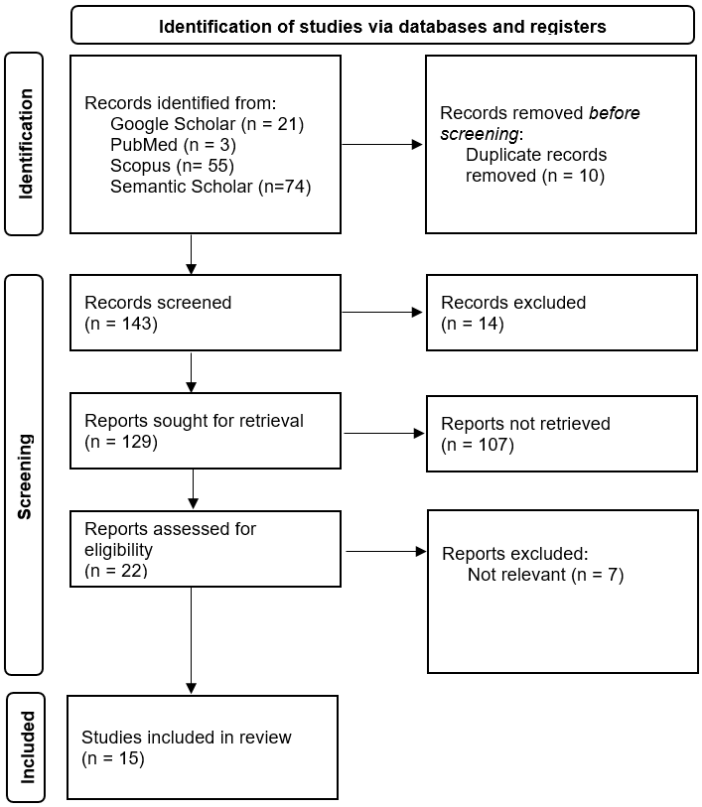

Fig. 1. PRISMA flow diagram for the article selection process.

3. State of the art

This section provides a structured overview of the key findings from the literature identified through the previously described methodology. We present a synthesis of the datasets commonly used in skin lesion classification, the evaluation metrics employed to assess model performance, and a detailed comparative analysis of the classification models proposed in the reviewed studies.

3.1. Skin disease datasets

Datasets are essential for the development of effective machine learning models, particularly in dermatology. They offer a collection of images and metadata that enable algorithms to learn how to identify and diagnose skin diseases. The table 1 provides an overview of key datasets used in scientific literature, highlighting their size, number of disease classes, and types of pathologies represented.

The PAD-UFES-20 dataset [2] contains a variety of skin conditions, including Basal Cell Carcinoma (BCC), Squamous Cell Carcinoma (SCC), Actinic Keratosis (ACK), Seborrheic Keratosis (SEK), Bowen's Disease (BOD), Melanoma (MEL), and Nevus (NEV), with Bowen's Disease classified as a variant of SCC.

The MSLD dataset [3] features diseases such as Mpox (MKP), Chickenpox (CHP), Cowpox (CWP), Measles (MSL), Hand, Foot and Mouth Disease (HFMD), and a healthy class (HEALTHY). An overview is presented in Figure 2.

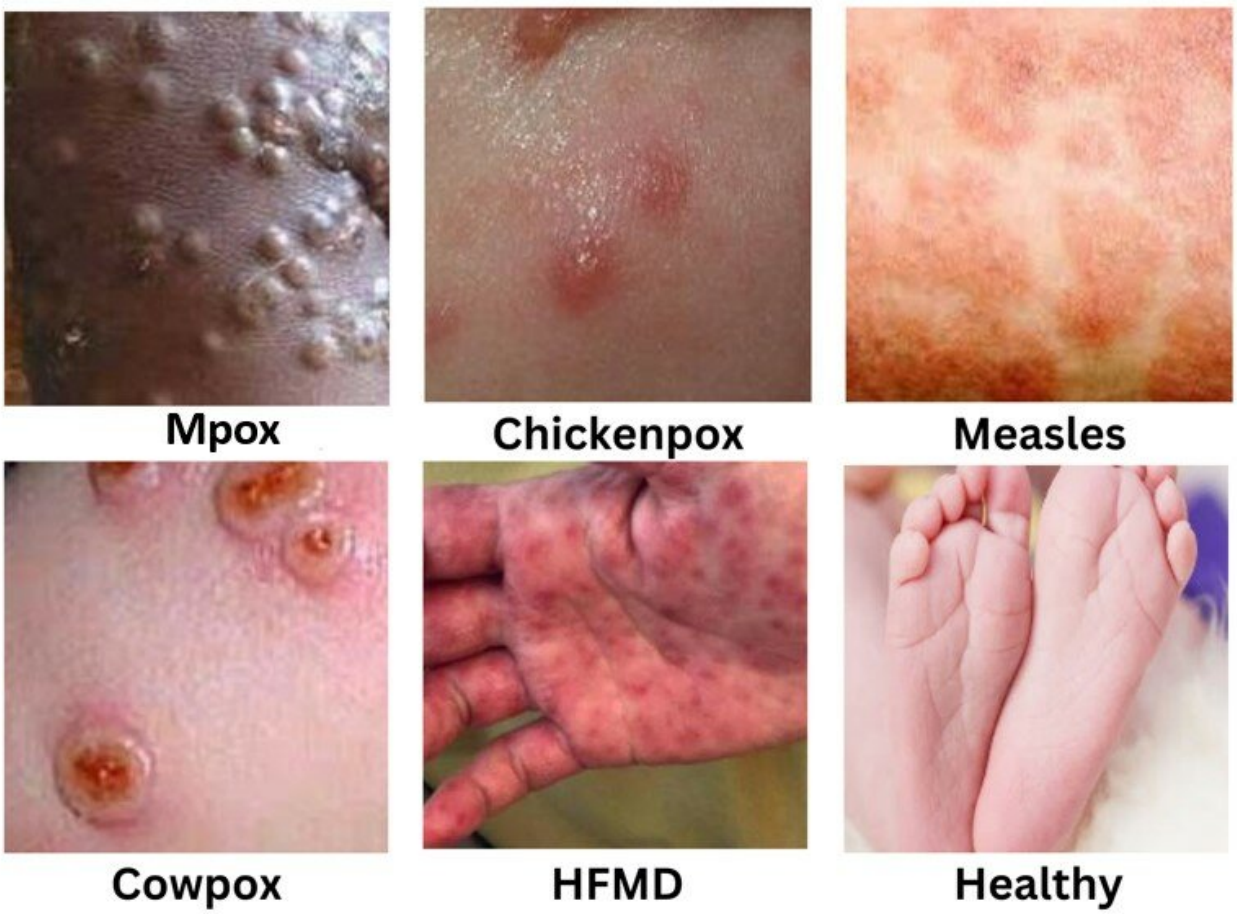

Fig. 2. Overview of Mpox (MKP), Chickenpox (CHP), Cowpox (CWP), Measles (MSL), Hand, Foot and Mouth Disease (HFMD), and a healthy class (HEALTHY) [3].

Table 1
Datasets summary.

Dataset	Class	Number of image
PAD-UFES-20 [2]	6	2298
MSLD [3, 18]	6	755
Department of Dermatology at the Third Xiangya Hospital of Central South University [4]	6	1366
National Skin Disease Database (NSDD) [5]	5	16313
SD-198 [6]	198	6584
Dataset of Dr. Gerbi Medium Clinic in Jimma, Ethiopia [9]	4	407
Dataset provided from photographs collected prospectively in Côte d'Ivoire and Ghana [10]	5	1709
PH2 [7]	3	200
HAM10000 [8]	7	10015
ISIC2016 [11]	2	900
ISIC2017 [12]	3	2000
ISIC2018 [13, 14]	7	10015
ISIC2019 [12, 13, 15]	8	25331
ISIC2020 [16]	2	33126
[17]	7	3406

The dataset from the Department of Dermatology at the Third Xiangya Hospital of Central South University [4] includes images of Melasma (ML), Naevus Fusco-caeruleus Zygomaticus (NZ), Freckles (FC), Cafe-au-lait Spots (CS), Nevus of Ota (NO), and Lentigo Simplex (LS).

The National Skin Disease Database (NSDD) [5] includes conditions such as Atopic Dermatitis (AD), Mycosis Fungoides (MF), Impetigo (IM), Herpes Simplex, and Kaposi Varicelliform Eruption.

The SD-198 dataset [6] includes images of Eczema (ECZ), Acne (ACN), and various Cancerous Conditions (CNC).

The PH2 dataset [7] consists of 200 dermoscopic images, with three categories: Common Nevus, Atypical Nevus, and Melanoma. This dataset is particularly useful for distinguishing between different types of melanocytic lesions.

The HAM10000 dataset [8] includes 10,015 dermoscopic images encompassing a diverse range of pigmented lesions such as Actinic Keratosis (AKIEC), Basal Cell Carcinoma (BCC), Benign Keratosis (BKL), Dermatofibroma (DF), Melanoma (MEL), Melanocytic Nevus (NV), and Vascular Lesions (VASC).

The dataset of Dr. Gerbi Medium Clinic in Jimma, Ethiopia [9] includes images of Tinea Pedis (TP), Tinea Capitis (TC), Tinea Corporis (TCo), and Tinea Unguium (TU).

The dataset from photographs collected prospectively in Côte d'Ivoire and Ghana [10] includes Buruli Ulcer (BU), Leprosy (LEP), Mycetoma (MYC), Scabies (SCA), and Yaws (YAW).

The ISIC2016 dataset [11] comprises 900 dermoscopic images categorized into benign and malignant classes, serving as a foundational dataset for early studies in skin disease classification.

The ISIC2017 dataset [12] contains 2,000 dermoscopic images classified into three categories: Melanoma (malignant skin tumor derived from melanocytes), Nevus (benign skin tumor derived from melanocytes), and Seborrheic Keratosis (benign tumor derived from keratinocytes).

The ISIC2018 dataset [13, 14] contains seven classes of skin lesions, including Melanoma (MEL), Melanocytic Nevus (NV), Basal Cell Carcinoma (BCC), Actinic Keratosis/Bowen's Disease (AKIEC), Benign Keratosis (BKL), Dermatofibroma (DF), and Vascular Lesion (VASC). An overview is presented in Figure 3.

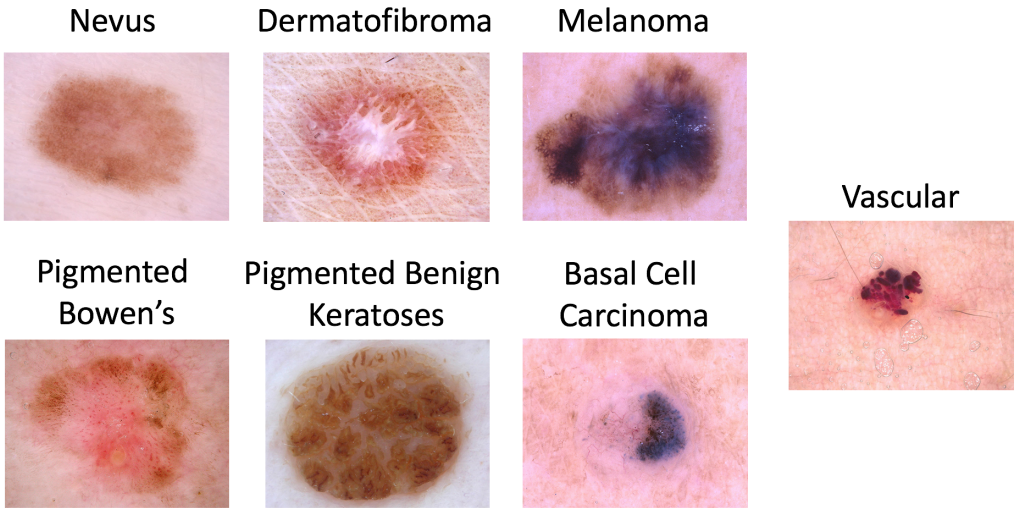

Fig. 3. Overview of Melanoma (MEL), Melanocytic Nevus (NV), Basal Cell Carcinoma (BCC), Actinic Keratosis/Bowen's Disease (AKIEC), Benign Keratosis (BKL), Dermatofibroma (DF), and Vascular Lesion (VASC) [14].

The ISIC2019 dataset [12, 13, 15] covers a wide spectrum of skin lesions, including Melanoma (MEL), Melanocytic Nevus (NV), Basal Cell Carcinoma (BCC), Actinic Keratosis (AK), Benign Keratosis (BK), Dermatofibroma (DF), Vascular Lesion (VL), and Squamous Cell Carcinoma (SCC).

The ISIC2020 dataset [16] contains 33,126 dermoscopic images of unique benign and malignant skin lesions from over 2,000 patients. This dataset is widely used for classification and segmentation tasks.

Finally, the dataset highlighted by Jessica et al. [17] consists of seven conditions: Acne (ACN), Varicella (CHP), Eczema (ECZ), Pityriasis Rosea (PR), Psoriasis (PSO), Vitiligo (VIT), and Tinea Corporis (TC). An overview is presented in Figure 4.

The diverse range of datasets available for dermatology research plays a critical role in the advancement of accurate and reliable diagnostic models. These datasets allows researchers to train and evaluate algorithms on various conditions, leading to improvements in the field of dermatological diagnosis and treatment. More detailed information on datasets is shown in Table 1.

3.2. Performance evaluation metrics

The evaluation of skin disease classification models was performed using standard metrics, where TP, TN, FP, and FN represent true positives, true negatives, false positives, and false negatives, respectively.

Accuracy is the most intuitive metric and measures the overall correctness of the model's predictions. It is defined as the ratio of correctly classified instances (both positive and negative) to the total number of instances:

$$Accuracy = \frac{TP + TN}{TP + TN + FP + FN} \quad (1)$$

Precision is a measure of how many instances predicted as positive are truly positive:

$$Precision = \frac{TP}{TP + FP} \quad (2)$$

Recall, also known as Sensitivity or True Positive Rate (TPR), measures the ability of the model to correctly identify all relevant instances:

$$Recall = \frac{TP}{TP + FN} \quad (3)$$

Fig. 4. Overview of Acne (ACN), Varicella (CHP), Eczema (ECZ), Pityriasis Rosea (PR), Psoriasis (PSO), Vitiligo (VIT), and Tinea Corporis (TC) [17].

The F1-score is the harmonic mean of Precision and Recall, providing a balanced measure that considers both false positives and false negatives:

$$F1 - score = 2 \cdot \frac{\text{Precision} \cdot \text{Recall}}{\text{Precision} + \text{Recall}} \quad (4)$$

3.3. Classification models

The classification of skin diseases using machine learning has attracted considerable interest in dermatology, with numerous models developed based on different datasets, architectures, and computational constraints. Although the reported performance varies widely, clear trends emerge regarding the relationship between model complexity, dataset diversity, and diagnostic accuracy.

ResNet-based models have been widely explored due to their strong feature extraction capability and efficient gradient propagation. In [19], ResNet-18 achieved 74.27% accuracy on a combined dataset (MSLD and PAD-UFES-20) and slightly higher performance (74.62%) on PAD-UFES-20 alone, suggesting that dataset heterogeneity can introduce noise that limits its generalization. On the ISIC dataset, ResNet-18 demonstrated remarkable performance (98.6%) in [20], outperforming VGG-16 and DenseNet201, which achieved 97.50% and 96.61% respectively. This result highlights ResNet's superior ability to handle texture and color variations, particularly in high-quality, well-annotated datasets. However, deeper architectures like ResNet-152 [21] showed diminishing returns, achieving 75.30% precision and 71.71% recall, indicating that beyond a certain depth, overfitting and computational cost may outweigh the benefits. To address interpretability challenges, a hybrid ResNet-50 with Radial Basis Function (RBF) networks [22] achieved balanced accuracy across ISIC2016 and ISIC2017, demonstrating how hybridization can enhance explainability without significant performance loss.

VGG-based models, though historically influential, tend to underperform on complex and imbalanced datasets due to their limited representational power. In [10], VGG-16 was outperformed by ResNet-50

(82.22% vs. 84.63%), underscoring the importance of residual connections for stable optimization. On ISIC2017, VGG-16 reached only 70.09% accuracy [22], confirming that its deeper layers struggle to capture fine-grained dermatological patterns compared to more modern architectures.

Lightweight models such as MobileNet have proven advantageous in resource-constrained environments. In [17], MobileNet achieved 94.1% accuracy across seven skin conditions, while in [4], its accuracy dropped to 70.39% on hyperpigmented diseases. This contrast suggests that MobileNet's performance depends heavily on the dataset's visual diversity and illumination consistency—critical factors for darker skin tones, where contrast is typically lower.

DenseNet architectures, known for feature reuse, produced mixed results. DenseNet201 achieved a precision of 73.28% on a custom dataset [4], comparable to VGG19. This suggests that feature redundancy may not necessarily enhance performance when data diversity or annotation quality is limited.

In contrast, EfficientNet and hybrid approaches have demonstrated strong adaptability. In [23], EfficientNetV2-B0 achieved an F1-score of 85.8% on ISIC2019, providing an optimal trade-off between accuracy and computational efficiency. Ensemble and hybrid frameworks further improved results: combining EfficientNet-V2 and Swin Transformer [24] achieved 99.10% F1-score on ISIC2018, while integrating ResNet-50 and EfficientNet with Unet3+ for joint segmentation-classification [25] yielded a recall of 96.45% and an F1-score of 98.78%. These findings confirm that multi-scale feature fusion and ensemble learning significantly enhance robustness and generalization across heterogeneous datasets.

Finally, bio-inspired optimization techniques have emerged as promising alternatives. In [26], a CNN trained with the Grey Wolf Optimization (GWO) algorithm achieved 95.11% accuracy and 96.16% F1-score on HAM10000, illustrating how adaptive optimization can improve convergence and classification balance, especially for imbalanced datasets.

In summary, ResNet and EfficientNet families dominate current research due to their strong trade-offs between depth, efficiency, and generalization, while lightweight models like MobileNet remain essential for real-time and mobile diagnostics. However, performance disparities across datasets underscore the ongoing challenge of achieving fairness and reliability, particularly for darker skin tones. A detailed synthesis of the models and their performance metrics is provided in Table 2.

Table 2
Performance of different classification model on various skin lesion datasets.

Method	Dataset	Class	Precision	Recall	F1-score	Accuracy
ResNet-18 [19]	MSLD + PAD-UFES-20	7	76.78	71.40	73.63	74.27
ResNet-18 [19]	PAD-UFES-20	6	75.17	62.77	65.90	74.62
VGG19 [4]	Dept. of Dermatology, Third Xiangya Hospital	–	73.28	–	–	–
DenseNet201 [4]	Dept. of Dermatology, Third Xiangya Hospital	–	73.28	–	–	–

Table 2 (continued)

Method	Dataset	Class	Precision	Recall	F1-score	Accuracy
MobileNet [4]	–	–	–	–	–	70.39
MobileNet [17]	–	–	–	–	–	94.10
ResNet-18 [20]	ISIC (5 classes)	5	97.58	97.42	97.37	98.6
VGG-16 [20]	ISIC (5 classes)	5	97.48	97.42	97.41	97.50
DenseNet201 [20]	ISIC (5 classes)	5	97.13	97.09	97.10	96.61
EfficientNetv2-B0 [23]	ISIC2019	8	85.80	98.00	85.80	85.50
ResNet-50 [10]	Côte d'Ivoire and Ghana dataset	–	84.63	–	–	–
VGG-16 [10]	Côte d'Ivoire and Ghana dataset	5	–	82.22	–	–
EfficientNet-V2 + Swin-Transformer (Ensemble DL Model) [24]	ISIC2018	7	–	99.27	–	99.10
ResNet-152 [21]	ISIC2019	8	75.30	71.71	73.01	–
ResNet-50 [27]	ISIC2018	7	97.09	98.15	97.85	97.13
ResNet-50 [22]	ISIC2016	2	–	–	–	83.02
ResNet-50 [22]	ISIC2017	3	–	–	–	76.15
VGG-16 [22]	ISIC2016	2	–	–	–	79.54
VGG-16 [22]	ISIC2017	3	–	–	–	70.09
MFFDCNN-CTDC [25]	ISIC2017	3	–	96.45	96.54	98.78
MFFDCNN-CTDC [25]	HAM10000	7	–	86.58	89.05	98.89
GWO-CNN [26]	HAM10000	7	94.56	93.88	96.16	95.11

To provide a more intuitive understanding of the trade-offs between model performance and computational complexity, Figure 5 presents a comparative visualization of accuracy versus model complexity (measured in number of parameters and FLOPs) for architectures evaluated on ISIC benchmark datasets. This chart illustrates how ResNet-18 and EfficientNetV2-B0 achieve competitive accuracy with relatively lower computational overhead compared to deeper architectures such as ResNet-152 and DenseNet201. The visualization confirms that increasing model depth does not necessarily guarantee proportional performance gains, particularly when datasets are limited or imbalanced. Lightweight models such as MobileNet and EfficientNet emerge as optimal choices for deployment in resource-constrained settings, balancing diagnostic accuracy with inference efficiency.

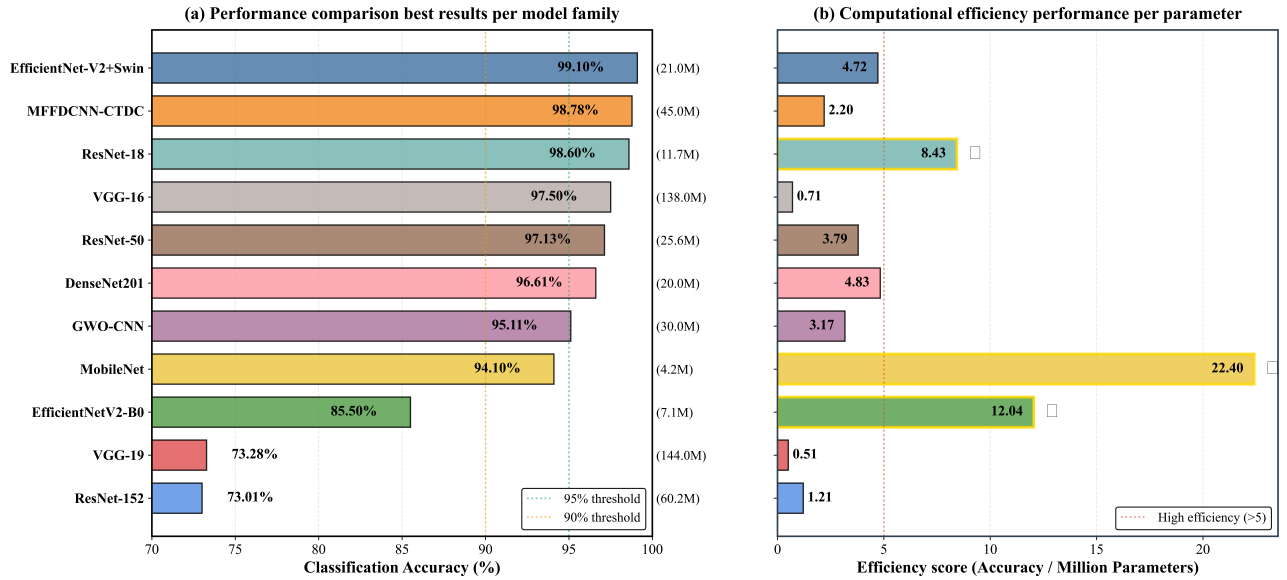

Fig. 5. Accuracy versus model complexity for deep learning models in skin disease classification.

4. Discussion

Model selection for skin disease classification depends on the target device, dataset diversity, and computational constraints. Traditional architectures (ResNet, VGG, DenseNet) remain robust baselines, but newer models (EfficientNet, MobileNet) introduce optimizations that improve the accuracy-cost trade-off through depthwise separable convolutions and compound scaling. These advances enable real-time deployment in resource-limited settings, critical for mobile diagnostics in low-resource environments. Performance differences between architectures stem from design principles and dataset characteristics. ResNet-50 outperformed VGG-16 for tropical diseases [10], leveraging residual connections for gradient flow in heterogeneous African clinical images. ResNet-18 excelled on ISIC [20], benefiting from high-resolution dermoscopic images with clear boundaries. DenseNet201 showed feature redundancy on homogeneous datasets, highlighting the importance of architecture-data fit importance. EfficientNetV2-B0 [23] maintains high precision on diverse datasets like ISIC-2019 despite resolution variability. Hybrid models (e.g., ResNet-50 with RBF networks [22]) improve interpretability through clinician-understandable decision boundaries, advancing explainable AI in dermatology. Ensemble and multi-scale fusion methods expand accuracy frontiers. Combining architectures (EfficientNet+Swin Transformer [24, 25]) captures complementary spatial and semantic features. Bio-inspired optimization (Grey Wolf [26]) dynamically refines parameters, improving adaptability across diverse illumination, texture, and pigmentation conditions. Despite progress, dataset diversity remains limited, particularly for dark skin tones. Public datasets (ISIC, PAD-UFES-20) overrepresent lighter skin types [10], biasing feature extraction and reducing accuracy for darker phototypes. This affects generalization and raises ethical deployment concerns, necessitating balanced datasets and fairness-aware training. Optimizing lightweight models for mobile deployment is critical for accessibility in sub-Saharan Africa, where dermatological expertise is scarce.

5. Limitations

This review has several limitations that should be acknowledged. First, potential publication bias may exist, as we included only peer-reviewed studies and preprints available through major academic databases, potentially excluding unpub-

lished negative results or studies in low-impact journals. Second, our language restriction to English-language publications may have excluded relevant research published in other languages, particularly regional studies from non-English-speaking African, Asian, or Latin American countries where skin disease prevalence differs significantly. Third, the heterogeneity of evaluation protocols across studies complicates direct comparisons; different train-test splits, data augmentation strategies, and validation procedures introduce variability that may affect performance benchmarking. Fourth, limited Fitzpatrick skin type reporting in most studies prevents systematic analysis of model fairness across diverse populations. Finally, the rapid evolution of AI architectures means that some models analyzed here may become outdated quickly, necessitating continuous review updates.

6. Conclusion

This study provided a systematic and comparative analysis of recent advances in AI-driven skin disease classification, emphasizing the performance, efficiency, and fairness of deep learning architectures between 2023 and 2025. Traditional architectures such as ResNet, VGG, and DenseNet have consistently shown strong performance in controlled environments but often struggle with scalability and dataset diversity. Moreover, the rapid evolution of AI models raises legitimate concerns about the obsolescence of earlier architectures, underscoring the need for continuous innovation to maintain clinical relevance.

Newer architectures such as EfficientNet, MobileNet, and hybrid CNN–Transformer approaches have emerged to address these limitations by introducing improved feature scaling, reduced parameter complexity, and enhanced adaptability to heterogeneous data. Unlike previous reviews focusing on single datasets or architectures, this work contributes a comprehensive comparative synthesis across multiple studies, supported by a standardized evaluation grid covering accuracy, precision, recall, and F1-score. By mapping these metrics to dataset characteristics (size, image quality, Fitzpatrick distribution), our review highlights how data imbalance, particularly the underrepresentation of dark skin tones, continues to affect model reliability and generalization.

We also provide a cross-architectural comparison showing that lightweight models (e.g., MobileNet, EfficientNet) perform competitively in low-resource environments without compromising diagnostic accuracy. This finding has practical implications for the deployment of mobile dermatological applications, especially in African and low-income regions where computational resources and dermatological expertise are limited.

In addition, emerging research trends such as bio-inspired optimization methods (e.g., Grey Wolf Optimization) and explainable AI frameworks aim to improve interpretability, efficiency, and fairness. However, a major challenge persists: the scarcity of diverse and demographically balanced datasets, particularly those representing darker skin tones. Addressing this issue remains crucial to ensuring reliable, ethical, and inclusive AI-based dermatological diagnostics.

Future research should focus on the following directions:

- **Building balanced and inclusive datasets**: Prioritize prospective data collection in underrepresented regions (e.g., sub-Saharan Africa) covering all Fitzpatrick types (I-VI). Establish partnerships with local dermatology clinics in countries like Burkina Faso, Ghana, and Kenya to capture diverse skin conditions under varying lighting and clinical settings;

- **Integrating fairness-aware AI strategies**: Implement data resampling techniques (e.g., SMOTE for minority class oversampling, synthetic image generation via GANs for

underrepresented skin tones), adversarial de-biasing methods (training models to learn skin-tone-invariant representations), and fairness constraints during optimization (ensuring equal accuracy, sensitivity, and specificity across demographic groups);

- **Deploying mobile diagnostic tools**: Develop lightweight models optimized for offline inference on Android devices, leveraging TensorFlow Lite or ONNX Runtime. Pilot programs in rural African clinics can validate real-world performance while addressing connectivity and hardware limitations;

- **Continuous model updating**: Establish systematic protocols for retraining models as new architectures and datasets emerge, preventing technological obsolescence while maintaining clinical relevance and regulatory compliance.

By addressing these priorities, AI-driven skin disease classification systems can become more equitable, transparent, and clinically applicable worldwide.

Abbreviations used

- AI: Artificial Intelligence
- CNN: Convolutional Neural Network
- cNTDs: Cutaneous Neglected Tropical Diseases
- FN: False Negative
- FP: False Positive
- GWO: Grey Wolf Optimization
- ISIC: International Skin Imaging Collaboration
- NSDD: National Skin Disease Database
- PRISMA: Preferred Reporting Items for Systematic Reviews and Meta-Analyses
- RBF: Radial Basis Function
- SMOTE: Synthetic Minority Oversampling Technique
- TN: True Negative
- TP: True Positive
- WHO: World Health Organization

References

[1] WHO: WHO's first global meeting on skin NTDs calls for greater efforts to address their burden (2023). https://www.who.int/newsitem/31-03-2023-who-first-global-meeting-on-skin-ntds-calls-for-greater-efforts-to-address-their-burden

[2] A.G.C. Pacheco, G.R. Lima, A.S. Salomão, B. Krohling, I.P. Biral, G.G. de Angelo, F.C.R. Alves Jr *et al.*, PAD-UFES-20: a skin lesion dataset composed of patient data and clinical images collected from smartphones, Mendeley Data, V1 (2020). https://doi.org/10.17632/zr7vgbcyr2.1

[3] N.A. Shams, A.Md. Tazuddin, T. Jahan, J. Paul, S.M.S. Sani, N. Noor, A.N Asma, T. Hasan, *A web-based Mpox skin lesion detection system using state-of-the-art deep learning models considering racial diversity*, Biomedical Signal Processing and Control 98 (2024) 106742 (2024). https://doi.org/10.48550/arXiv.2306.14169

[4] J. Lu, X. Tong, H. Wu, Y. Liu, H. Ouyang, Q. Zeng, *Image classification and auxiliary diagnosis system for hyperpigmented skin diseases based on deep learning*, Heliyon 9(9) (2023) e20186. https://doi.org/10.1016/j.heliyon.2023.e20186

[5] Y. Yanagisawa, K. Shido, K. Kojima, K. Yamasaki, *Convolutional neural network-based skin image segmentation model to improve classification of skin diseases in conventional and non-standardized picture images*, Journal of Dermatological Science, Elsevier (2023). https://doi.org/10.1016/j.jdermsci.2023.01.005

[6] X. Sun, J. Yang, M. Sun, K. Wang, A benchmark for automatic visual classification of clinical skin disease images. In "B. Leibe, J. Matas, N. Sebe, M. Welling (eds.) Computer Vision – ECCV 2016, Lecture Notes in Computer Science 9910, Springer, Cham (2016)". https://doi.org/10.1007/978-3-319-46466-4_13

[7] T. Mendonca, P.M. Ferreira, J.S. Marques, A.R. Marcal, J. Rozeira, PH² - a dermoscopic image database for research and benchmarking, Annual International Conference of the IEEE Engineering in Medicine and Biology Society (EMBC) (2013) 5437-5440. https://doi.org/10.1109/EMBC.2013.6610779

[8] P. Tschandl, The HAM10000 dataset, a large collection of multi-source dermatoscopic images of common pigmented skin lesions, Harvard Dataverse (2018), Version V4. https://doi.org/10.7910/DVN/DBW86T

[9] T.D. Nigat, *Fungal skin disease classification using the convolutional neural network*, Journal of Healthcare Engineering (2023) 6370416. https://doi.org/10.1155/2023/6370416

[10] R. Yotsu, Z. Ding, J. Hamm, R. Blanton, Deep learning for AI-based diagnosis of skin-related neglected tropical diseases: A pilot study (2023). https://doi.org/10.1371/journal.pntd.0011230

[11] D. Gutman, N.C.F. Codella, E. Celebi, B. Helba, M. Marchetti, N. Mishra, A. Halpern, 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 (2016). https://doi.org/10.48550/arXiv.1605.01397

[12] N.C.F. Codella, D. Gutman, M.E. Celebi, B. Helba, M.A. Marchetti, S.W. Dusza, A. Kalloo, K. Liopyris, N. Mishra, H. Kittler, A. Halpern, 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), arXiv (2017). https://doi.org/10.48550/arXiv.1710.05006

[13] P. Tschandl, C. Rosendahl, H. Kittler, *The HAM10000 dataset, a large collection of multi-source dermatoscopic images of common pigmented skin lesions*, Scientific Data 5 (2018) 180161. https://doi.org/10.1038/sdata.2018.161

[14] N. Codella, V. Rotemberg, P. Tschandl, M.E. Celebi, S. Dusza, D. Gutman, B. Helba *et al.*, Skin lesion analysis toward melanoma detection 2018: A challenge hosted by the International Skin Imaging Collaboration (ISIC), arXiv (2018). https://doi.org/10.48550/arXiv.1902.03368

[15] C. Hernández-Pérez, M.Combalia, S. Podlipnik *et al*, *BCN20000: Dermoscopic lesions in the wild*, Scientific Data 11(1) (2024) 641. https://doi.org/10.1038/s41597-024-03387-w

[16] V. Rotemberg, N. Kurtansky, B. Betz-Stablein, L. Caffery, E. Chousakos, N. Codella, M. Combalia *et al.*, *A patient-centric dataset of images and metadata for identifying melanomas using clinical context*, Sci Data 8 (2021) 34. https://doi.org/10.1038/s41597-021-00815-z

[17] J.S. Velasco, J.V. Catipon, E.G. Monilar *et al.*, Classification of skin disease using transfer learning in convolutional neural networks, arXiv(2023). https://doi.org/10.48550/arXiv.2304.02852

[18] S.N. Ali, M.T. Ahmed, J. Paul *et al.*, Monkeypox Skin Lesion Detection using deep learning models: A preliminary feasibility study, arXiv (2022). https://doi.org/10.48550/arXiv.2207.03342

[19] I. Oztel, G. Yolcu Oztel, V.H. Sahin, Deep learning-based skin diseases classification using smartphones, Advanced Intelligent Systems, Wiley Online Library (2023). https://doi.org/10.1002/aisy.202300211

[20] F. Doğan, M. Aktaş, M.I. Gürsoy, Classification of skin diseases with different deep learning models and comparison of the performances of the models, Türk Doğa ve Fen Dergisi (2024). https://doi.org/10.46810/tdfd.1502471

[21] R. Agarwal, D. Godavarthi, *Skin disease classification using CNN algorithms*, EAI Endorsed Transactions on Pervasive Health and Technology 9 (2023) 1-8. https://doi.org/10.4108/eetpht.9.4039

[22] M.A. Ullah, T. Zia, Hybrid interpretable deep learning framework for skin cancer diagnosis: Integrating radial basis function networks with explainable AI, arXiv (2025). https://arxiv.org/abs/2501.14885

[23] E.H. Kırğıl, Ç.B. Erdaş*Enhancing skin disease diagnosis through deep learning: A comprehensive study on dermoscopic image preprocessing and classification*, International Journal of Imaging Systems and Technology 34 (2024) e23148. https://doi.org/10.1002/ima.23148

[24] K. Shehzad, T. Zhenhua, S. Shoukat *et al.*, *A deep-ensemble-learning-based approach for skin cancer diagnosis*, Electronics 12(6) 1342. https://doi.org/10.3390/electronics12061342.

[25] U M. Prakash, S. Iniyan, A.K. Dutta *et al.*, *Multi-scale feature fusion of deep convolutional neural networks on cancerous tumor detection and classification using biomedical images*, Scientific Reports 15 (2025) 1105. https://doi.org/10.1038/s41598-024-84949-1

[26] F. Mazhar, N. Aslam, A. Naeem, H. Ahmad, M. Fuzail, M. Imran, *Enhanced diagnosis of skin cancer from dermoscopic images using alignment optimized convolutional neural networks and Grey Wolf Optimization*, Journal of Computing Theories and Applications 2(3) (2025) 368-382. https://doi.org/10.62411/jcta.11954

[27] N.M. Rashad, N.M. Abdelnapi, A.F. Seddik *et al.*, *Automating skin cancer screening: a deep learning*, Journal of Engineering and Applied Science 72(6) (2025) 1-18. https://doi.org/10.1186/s44147-024-00573-w