Research Square

Preprints are preliminary reports that have not undergone peer review.
They should not be considered conclusive, used to inform clinical practice,
or referenced by the media as validated information.

Melanoma Detection Using Deep Convolutional Neural Networks: Architectures, Datasets, and Performance

Dr Kazi Abdul Mannan

drkaziabdulmannan@gmail.com

Center for Academic & Professional Career Development and Research (CAPCDR)
https://orcid.org/0000-0002-7123-132X

Mashrafi Ahmed

Shanto-Mariam University of Creative Technology https://orcid.org/0009-0001-6295-8113

Research Article

Keywords: Melanoma detection, Convolutional Neural Network, Dermoscopy, Transfer learning, Skin lesion classification, Deep learning

Posted Date: August 3rd, 2026

DOI: https://doi.org/10.21203/rs.3.rs-10547895/v1

License: This work is licensed under a Creative Commons Attribution 4.0 International License.
Read Full License

Additional Declarations: The authors declare no competing interests.

Abstract

Melanoma is the most lethal form of skin cancer, with survival rates declining sharply upon metastasis. Early and accurate detection is critical for improving patient outcomes. Traditional diagnostic methods depend on clinical expertise and are prone to inter-observer variability. Deep Convolutional Neural Networks (CNNs) have emerged as a powerful tool for automated melanoma detection from dermoscopic images, offering high accuracy, scalability, and reproducibility. This study presents a systematic review of CNN-based approaches for melanoma classification, examining key architectures including ResNet, EfficientNet, DenseNet, and NASNet, alongside benchmark datasets such as HAM10000 and the ISIC Archive. The review covers preprocessing pipelines, transfer learning strategies, data augmentation techniques, and model evaluation metrics. Leading models have achieved accuracy rates of up to 95.2% and AUC scores of 95.75%. Challenges including class imbalance, dataset heterogeneity, and model interpretability are discussed. Future directions involving federated learning, hybrid CNN-Transformer models, and mobile deployment are also explored.

1. Introduction

Melanoma is a malignant neoplasm arising from melanocytes, the pigment-producing cells of the skin. Although it represents only approximately 1% of all skin cancer diagnoses, melanoma accounts for the overwhelming majority of skin cancer-related mortality [1]. The global incidence of melanoma has risen steadily over recent decades, with over 100,000 new cases reported annually in the United States alone, and significantly higher rates in Australia, New Zealand, and parts of Northern Europe [2]. The five-year survival rate for localized melanoma exceeds 98%; however, this rate drops to approximately 25% once distant metastasis has occurred [3]. Ultraviolet (UV) radiation - from both natural sunlight and artificial tanning devices - is the primary environmental risk factor, with additional contributions from genetic predisposition, fair skin phenotype, and immunosuppression [4]. Conventional diagnosis relies on clinical examination, dermoscopy, and histopathological biopsy, all of which are resource-intensive and subject to inter-clinician variability. The integration of artificial intelligence into dermatological imaging represents a transformative opportunity to standardize and accelerate melanoma diagnosis [5].

Despite advances in dermatological imaging, melanoma remains critically underdiagnosed in its early stages. The diagnostic accuracy of unaided visual inspection by general practitioners is reported at approximately 60–70%, while even experienced dermatologists demonstrate significant variability [6]. Dermoscopy improves sensitivity but requires specialized training and remains subjective. Furthermore, access to expert dermatologists is severely limited in rural, low-income, and underserved regions, creating significant health disparities in early detection [7]. Existing computer-aided diagnosis (CAD) systems based on traditional machine learning - relying on hand-crafted features such as color histograms, texture descriptors, and shape parameters - have failed to achieve the accuracy and generalizability required for clinical deployment [8]. Deep learning, particularly CNNs, has demonstrated the potential to overcome these limitations through automatic hierarchical feature learning from raw pixel data. However, critical challenges persist: severe class imbalance in benchmark datasets, poor

cross-dataset generalization, lack of demographic diversity, and the "black box" nature of deep models that undermines clinical trust and regulatory compliance [9]. A systematic understanding of these challenges is essential for advancing reliable CNN-based melanoma detection.

The rationale for this research lies in the urgent clinical need for accurate, scalable, and equitable melanoma detection tools. While numerous individual CNN-based studies have been published, there is a lack of comprehensive, structured synthesis examining architectures, datasets, methodologies, and performance outcomes in a unified framework [10]. Understanding the comparative strengths and limitations of existing approaches is essential for guiding future model development and clinical translation. This review is motivated by the observation that leading CNN models have achieved dermatologist-level performance in controlled settings, yet remain largely absent from routine clinical practice - a gap attributable to interpretability concerns, dataset bias, and insufficient validation across diverse populations [11]. A rigorous analysis of the current state of the art will provide actionable insights for researchers, clinicians, and policymakers.

This study aims to:

- Systematically review CNN architectures applied to melanoma detection from dermoscopic images, including ResNet, EfficientNet, DenseNet, InceptionNet, and NASNet.
- Identify and compare benchmark datasets used in the field, including HAM10000, ISIC 2016–2020, and PH2.
- Analyze preprocessing, segmentation, data augmentation, and transfer learning methodologies employed in leading studies.
- Evaluate and compare reported performance metrics - accuracy, AUC, sensitivity, specificity, and F1 score - across key studies.
- Identify persistent challenges including class imbalance, generalization failure, demographic bias, and interpretability limitations.
- Outline future research directions including federated learning, vision transformers, and mobile deployment strategies.

This research holds significant implications for both the scientific community and clinical practice. By providing a structured synthesis of CNN-based melanoma detection approaches, it enables researchers to identify methodological gaps and build upon the most effective strategies. For clinicians, it offers a clear understanding of the current capabilities and limitations of AI-assisted dermoscopy, facilitating informed adoption decisions. For policymakers and healthcare administrators, the review highlights the potential for deep learning tools to address diagnostic inequities in underserved regions through scalable, low-cost screening platforms [12]. Furthermore, the identification of bias and generalization challenges underscores the need for diverse, representative datasets - a policy-relevant finding with direct implications for equitable AI development. This work contributes to the broader goal of

responsible AI integration into healthcare, where accuracy, transparency, and fairness are equally essential.

2. Literature Review

The application of machine learning to medical image analysis has a history spanning several decades, beginning with rule-based expert systems and evolving through traditional feature engineering approaches. Early computer-aided detection (CAD) systems for skin lesion analysis relied on manually crafted features - including ABCD rule parameters (asymmetry, border, color, diameter), texture descriptors such as local binary patterns (LBP), and color space histograms - fed into classifiers such as support vector machines (SVMs) and random forests [13]. While these methods demonstrated feasibility, their performance was constrained by the quality of hand-engineered features and their inability to generalize across imaging devices and patient populations.

The landmark publication of AlexNet in 2012 by Krizhevsky et al., which achieved unprecedented accuracy on the ImageNet Large Scale Visual Recognition Challenge (ILSVRC), catalyzed a

paradigm shift toward deep learning in computer vision [14]. CNNs demonstrated the ability to learn discriminative feature hierarchies directly from raw pixel data, eliminating the need for manual feature engineering. This breakthrough rapidly extended to medical imaging, with early applications in radiology (chest X-ray analysis, diabetic retinopathy screening) followed by dermatology. The 2017 landmark study by Esteva et al. in *Nature* demonstrated that a CNN trained on 129,450 clinical images classified skin cancer with dermatologist-level accuracy, establishing the field's foundational benchmark [15]. Subsequent ISIC challenges (2016–2024) provided standardized datasets and evaluation protocols, accelerating community-wide progress in dermoscopy-based melanoma detection.

2.1 CNN Architectures in Melanoma Detection

The evolution of CNN architectures has directly shaped the trajectory of melanoma detection research. VGGNet [16], introduced by Simonyan and Zisserman (2014), established deep uniform architectures using small 3×3 filters and became an early transfer learning backbone for dermoscopy classification due to its simplicity. However, its large parameter count (~ 138 million for VGG-16) limited scalability.

ResNet [17], introduced by He et al. (2016), resolved the vanishing gradient problem through residual skip connections, enabling training of networks exceeding 100 layers. In melanoma research, ResNet-101 has been extensively validated. A PMC study using the ISIC 2018 HAM10000 dataset reported ResNeXt - an extension of ResNet incorporating grouped convolutions - achieving 0.88 precision, 0.83 sensitivity, 0.85 F1 score, and 0.99 specificity across seven skin lesion classes [18]. SE-ResNet and SE-ResNeXt, which add channel-wise attention via squeeze-and-excitation blocks, demonstrated slightly lower precision in the same study, suggesting that architectural complexity does not always yield proportional gains.

InceptionNet (GoogLeNet) [19] introduced parallel multi-scale convolutions within inception modules, enabling efficient multi-resolution feature extraction. Mosa et al. (2022) reported InceptionV4 achieving 82.6% accuracy on an augmented dataset of over 21,000 images, outperforming ResNet-18 (79.5%) and DenseNet-121 (81.2%) in the same experimental setting [20]. DenseNet [21], which connects each layer to every subsequent layer, promotes feature reuse and gradient flow, making it well-suited to medical imaging tasks with limited data.

EfficientNet [22], proposed by Tan and Le (2019), introduced compound scaling - simultaneously scaling network width, depth, and input resolution using a fixed coefficient. EfficientNet variants (B0-B7) have dominated recent ISIC challenge leaderboards due to their superior accuracy-to-parameter ratio. NASNet [23], which employs neural architecture search to discover optimal network topologies, achieved 85.62% accuracy (NASNet Mobile) on a 2,637-image skin dataset in a 2024 study by Rahman et al. [24].

2.2 Transfer Learning and Data Augmentation

Given the scarcity of annotated dermoscopy images relative to natural image datasets, transfer learning from ImageNet-pretrained weights is the dominant training strategy across the literature [25]. Fine-tuning pretrained CNNs reduces the data requirement substantially while leveraging low-level feature representations (edges, textures) learned from millions of natural images. Studies consistently report that full fine-tuning outperforms feature extraction alone for dermoscopy tasks.

Data augmentation is universally employed to address class imbalance - a critical issue in melanoma datasets where benign nevi may constitute up to 67% of samples while melanoma represents only 11% [26]. Geometric transformations (rotation, flipping, zooming), color jittering, and elastic deformations are standard. Generative Adversarial Networks (GANs), including DCGAN and StyleGAN, have been applied to synthesize realistic melanoma images, with studies reporting improved CNN performance when GAN-augmented data is included in training [27].

2.3 Multi-Modal and Ensemble Approaches

Multi-modal learning - integrating dermoscopic images with clinical metadata (patient age, sex, lesion location) - has demonstrated consistent accuracy improvements over image-only models [28]. A study combining macroscopic and dermatoscopic image streams reported accuracy of 0.721, compared to 0.707 for dermatoscopy alone and 0.647 for macroscopy alone. Ensemble methods, combining predictions from multiple CNN models via probability averaging or majority voting, are employed by top ISIC challenge competitors, typically using 5–20 model ensembles [29].

2.4 Explainability

Gradient-weighted Class Activation Mapping (Grad-CAM) has emerged as the dominant explainability technique, generating heatmaps that highlight image regions driving model predictions [30]. LIME and SHAP provide complementary model-agnostic explanations. These tools are essential for clinical adoption, as regulatory frameworks (FDA, CE marking) increasingly require AI systems to provide interpretable outputs.

Despite significant progress, critical gaps persist in the literature. Most high-performing CNN models are validated exclusively on Western, fair-skinned populations, with HAM10000 and ISIC datasets offering limited representation of Fitzpatrick skin types IV-VI [31]. Cross-dataset generalization remains poorly studied, with most papers reporting performance on held-out splits of a single dataset rather than external validation cohorts. The integration of explainability techniques into the training pipeline - rather than as a post-hoc analysis - remains underdeveloped.

Furthermore, comparative studies systematically evaluating the same architecture across multiple datasets under identical preprocessing and augmentation conditions are rare, making it difficult to draw definitive conclusions about architectural superiority. Prospective clinical validation studies comparing CNN predictions to real-world dermatologist decisions under routine clinical conditions are virtually absent from the literature.

The literature demonstrates a clear trajectory from hand-crafted feature engineering toward end-to-end deep learning for melanoma detection, with CNNs achieving accuracy levels that rival experienced dermatologists in controlled benchmark settings. Transfer learning from ImageNet-pretrained architectures, combined with aggressive data augmentation and ensemble strategies, constitutes the current methodological consensus. ResNet, EfficientNet, and DenseNet variants dominate the literature, with EfficientNet emerging as the architecture of choice for recent high-performance systems. However, the field faces persistent challenges in dataset diversity, cross-institutional generalization, and clinical explainability that must be resolved before widespread deployment. The ISIC challenge series has provided invaluable standardization, but benchmark saturation - where models achieve near-perfect scores on held-out test sets while failing on real-

world data - is an emerging concern. Future progress will require not only architectural innovation but also concerted efforts in dataset curation, federated learning for privacy-preserving multi-institutional training, and rigorous prospective clinical trials that bridge the gap between research benchmarks and clinical reality.

3. Theoretical and Conceptual Framework

The theoretical underpinning of this research draws from three intersecting domains: computer vision theory, statistical learning theory, and medical decision theory.

From computer vision, the hierarchical feature learning paradigm of CNNs - where successive convolutional layers learn progressively abstract representations from low-level edges and textures to high-level semantic features - provides the mechanistic basis for dermoscopic image classification [32]. The universal approximation theorem establishes that sufficiently deep and wide neural networks can approximate any continuous function, justifying the expressiveness required for complex visual discrimination tasks such as melanoma classification [33].

Statistical learning theory provides the framework for understanding generalization - the ability of a model trained on a finite sample to perform well on unseen data. Key concepts include the bias-variance tradeoff, regularization theory (L1/L2 penalties, dropout as a form of ensemble approximation), and the role of dataset size in determining generalization bounds [34]. The severe class imbalance in melanoma datasets is addressed through the theoretical lens of cost-sensitive learning, where misclassification costs are asymmetric - false negatives (missed melanoma) carry far higher clinical cost than false positives [35].

Medical decision theory frames the clinical deployment context: a diagnostic AI system must be evaluated not only on accuracy but on sensitivity-specificity tradeoffs, calibration (the alignment of predicted probabilities with true outcome frequencies), and decision thresholds that reflect clinical risk tolerance [36]. The receiver operating characteristic (ROC) curve and area under the curve (AUC) are the theoretically preferred metrics for binary classification tasks with class imbalance, as they are invariant to decision threshold and class prevalence.

The conceptual framework of this study organizes the melanoma detection pipeline into five sequential stages:

Stage 1 - Data Acquisition

Dermoscopic images are acquired using standardized dermoscopes and uploaded to curated repositories (ISIC Archive, HAM10000).

Stage 2 - Preprocessing

Raw images undergo artifact removal (hair, bubbles), noise filtering, color normalization, and resizing to model-compatible dimensions.

Stage 3 - Feature Extraction via CNN

A pretrained CNN backbone (ResNet, EfficientNet, etc.) extracts hierarchical feature maps from preprocessed images. Transfer learning initializes weights from ImageNet pretraining.

Stage 4 - Classification

Extracted features are passed through fully connected layers with softmax/sigmoid activation to produce class probability scores (melanoma vs. non-melanoma, or

multi-class).

Stage 5 - Evaluation and Explainability

Model performance is assessed using accuracy, AUC, sensitivity, specificity, and F1 score. Grad-CAM visualizations are generated to highlight diagnostically relevant image regions.

Binary cross-entropy loss for melanoma classification:

$L = -[y \cdot \log(\hat{y}) + (1 - y) \cdot \log(1 - \hat{y})]$

Where $y \in \{0,1\}$ is the true label and $\hat{y}$ is the predicted probability of melanoma.

For class-imbalanced training, focal loss is applied:

$FL(p_t) = -\alpha_t \cdot (1 - p_t)^{\gamma} \cdot \log(p_t)$

Where γ is the focusing parameter (typically 2) and α_t is the class-balancing weight.

4. Methodology

This study adopts a qualitative, secondary data research design, employing a systematic literature review methodology to synthesize existing evidence on CNN-based melanoma detection. The review follows the Preferred Reporting Items for Systematic Reviews and Meta-Analyses (PRISMA) guidelines, ensuring transparency, reproducibility, and rigor in the evidence synthesis process [37].

The primary data sources include peer-reviewed journals indexed in PubMed, PubMed Central (PMC), IEEE Xplore, ScienceDirect, and Nature Publishing Group. Supplementary sources include preprint repositories (arXiv) for recent technical developments, and the ISIC challenge proceedings for benchmark performance data. Search queries combined terms including "melanoma detection," "deep learning," "convolutional neural network," "dermoscopy," "skin lesion classification," and "transfer learning," restricted to publications from 2016 to 2026.

Inclusion criteria required studies to: apply CNN-based deep learning to melanoma or skin

lesion classification; use dermoscopic image datasets; report quantitative performance metrics (accuracy, AUC, sensitivity, specificity, or F1 score); and be published in peer-reviewed venues or as preprints with subsequent peer-reviewed citation. Exclusion criteria eliminated studies using only traditional machine learning without deep learning components, studies not reporting quantitative outcomes, and duplicate publications.

A total of 14 primary studies were identified for detailed analysis, supplemented by foundational architecture papers (AlexNet, ResNet, EfficientNet, DenseNet, NASNet) and methodological references on transfer learning, data augmentation, and explainability. Data extraction was performed systematically, capturing: author(s), year, dataset, CNN architecture, preprocessing pipeline, augmentation strategy, training configuration, and reported performance metrics.

The qualitative synthesis focuses on identifying patterns, methodological trends, and performance determinants across studies, rather than quantitative meta-analysis - which is precluded by the heterogeneity of datasets, evaluation protocols, and reported metrics across the included literature.

4.1 Variable Selection

The primary independent variables in this review are: CNN architecture (ResNet, EfficientNet, DenseNet, InceptionNet, NASNet, VGGNet, MobileNet); dataset used (HAM10000, ISIC 2016–2020, PH2, Kaggle); preprocessing strategy (with/without artifact removal, normalization method); augmentation strategy (geometric, color, GAN-based); and training strategy (transfer learning vs. training from scratch; feature extraction vs. full fine-tuning).

The primary dependent variables are classification performance metrics: accuracy, AUC, sensitivity (recall), specificity, precision, and F1 score. Secondary dependent variables include model parameters count, computational efficiency (inference time, FLOPs), and interpretability (availability of Grad-CAM or equivalent visualization).

4.2 Cross Validation

Cross-validation is the primary strategy for unbiased performance estimation in the included studies. The most commonly employed approach is k-fold cross-validation (typically k = 5 or k = 10), where the dataset is partitioned into k equal folds; the model is trained on k-1 folds and validated on the remaining fold, with the process repeated k times and results averaged [38]. For the HAM10000 dataset, the official ISIC 2018 challenge split - with a predefined held-out test set - is widely used to enable fair comparison across studies. Stratified k-fold cross-validation is employed where class imbalance is severe, ensuring each fold maintains the original class distribution. Some studies additionally report performance on external validation sets (e.g., training on HAM10000, testing on PH2) to assess cross-dataset generalization, though this practice remains less common than single-dataset evaluation.

4.3 Model Evaluation

Model performance is evaluated using a comprehensive set of metrics appropriate for imbalanced binary and multi-class classification:

- **Accuracy**: Proportion of correctly classified samples; can be misleading under class imbalance.
- **AUC-ROC**: Area under the receiver operating characteristic curve; threshold-independent measure of discriminative ability; preferred metric for imbalanced datasets.
- **Sensitivity (Recall)**: True positive rate; clinically critical - measures the proportion of actual melanoma cases correctly identified.
- **Specificity**: True negative rate; measures the proportion of benign lesions correctly classified as benign.
- **Precision**: Proportion of predicted melanoma cases that are true melanoma.
- **F1 Score**: Harmonic mean of precision and recall; balances both metrics under class imbalance.

The confusion matrix provides the foundational data for computing all metrics. For multi-class tasks (HAM10000, 7 classes), macro-averaged and weighted-averaged metrics are reported.

4.4 Software

The primary deep learning frameworks used across the reviewed studies are TensorFlow (Google) and PyTorch (Meta), both of which provide high-level APIs for CNN construction, training, and evaluation. Keras (integrated with TensorFlow) is widely used for rapid prototyping. Python is the universal programming language. Data preprocessing employs OpenCV and PIL/Pillow. Scikit-learn provides evaluation metric computation. Grad-CAM implementations are available via the tf-explain and pytorch-grad-cam libraries. Experiments are conducted on NVIDIA GPU clusters (V100, A100) or Google Colab Pro environments.

4.5 Ethical Issue

The studies reviewed in this paper use publicly available, de-identified dermoscopic image datasets (HAM10000, ISIC Archive, PH2), which were collected under institutional ethics approval and patient consent at their originating institutions. No new patient data was collected for this review. A critical ethical concern identified across the literature is demographic bias: most datasets are skewed toward lighter skin tones, which may result in AI systems that underperform for patients with darker skin, raising equity and fairness concerns that must be addressed in future dataset curation and model validation efforts [39].

5. Results

The systematic review identified 14 primary studies reporting CNN-based melanoma or skin lesion classification performance, spanning the period 2021–2026. The following synthesis presents key findings organized by architecture, dataset, and performance outcome.

Table 5.1
Performance Summary by Study

Study	Dataset	Architecture	Accuracy	AUC	Sensitivity	Specificity	F1
Ahmad Hameed et al. (2023) [1]	HAM10000	CNN	95.2%	-	-	-	-
Md. Khairul Islam et al. (2021) [2]	HAM10000	CNN	90.93% (test)	-	-	-	-

Kalaycı, Serdar (2023) [5]	**ISIC 2019 & 2020**	**CNN**	**-**	**95.75%**	**-**	**-**	**-**
PMC Study (2021) [3]	HAM1000 0	ResNeXt	88% (avg)	0.99	0.83	0.99	0.85
Mosa et al. (2022) [6]	Augmente d	InceptionV 4	82.6%	-	-	-	-
Mosa et al. (2022) [6]	Augmente d	DenseNet-121	81.2%	-	-	-	-
Mosa et al. (2022) [6]	Augmente d	ResNet-18	79.5%	-	-	-	-
Rahman et al. (2024) [7]	2,637 images	NASNet Mobile	85.62%	-	-	-	-
Rahman et al. (2024) [7]	2,637 images	NASNet Large	83.98%	-	-	-	-
Pramanik & Chakrabort y (2021) [9]	Kaggle	CNN	87.58%	-	-	-	-
Raut et al. (2023) [8]	Dermosco py DB	CNN	90.1%	-	-	-	-
Maaz Ul Amin et al. (2024) [10]	HAM1000 0 + 3,672	CNN	84.27%	-	-	-	-
Oumoulylt e et al. (2023) [11]	ISIC	CNN	87%	-	-	-	-
Hermosilla et al. (2026) [12]	-	Hybrid CNN	High	-	-	-	-

5.1 Architecture Performance Trends

ResNeXt demonstrated the strongest overall performance profile in comparative studies, achieving the highest average precision (0.88) among ResNet family variants. The addition of squeeze-and-excitation (SE) blocks in SE-ResNet and SE-ResNeXt reduced precision by 0.05–0.07 points in the

PMC 2021 study, contrary to the expectation that attention mechanisms would improve performance - suggesting potential overfitting with added complexity on the relatively small HAM10000 dataset [3].

InceptionV4 consistently outperformed ResNet-18 and DenseNet-121 in the Mosa et al. (2022) multi-architecture comparison, achieving 82.6% vs. 81.2% and 79.5% respectively on augmented data [6]. This advantage is attributed to InceptionV4's multi-scale feature extraction capability, which captures both fine-grained texture patterns and broader structural features of dermoscopic lesions simultaneously.

NASNet Mobile achieved a higher accuracy (85.62%) than NASNet Large (83.98%) in the Rahman et al. (2024) study [7], a counterintuitive finding likely attributable to overfitting by the larger architecture on the relatively small 2,637-image dataset, underscoring the importance of dataset size in architecture selection.

5.2 Dataset and Augmentation Effects

Studies using the HAM10000 dataset with transfer learning and augmentation consistently reported test accuracies in the 84–95% range. The highest accuracy (95.2%) was reported by Ahmad Hameed et al. (2023), who employed an optimized CNN architecture with checkpoint-based training on HAM10000 [1]. The Kalaycı (2023) study, using the larger ISIC 2019 and 2020 datasets (25,331 and 33,126 images respectively), achieved the highest AUC of 95.75%, reflecting the benefit of larger training sets for discriminative model learning [5].

The Mosa et al. (2022) study demonstrated the impact of augmentation: by expanding the training set to over 21,000 images through augmentation from an initial smaller collection, InceptionV4 achieved 82.6% accuracy - a meaningful improvement over baseline performance reported on non-augmented subsets in the same experimental series [6].

5.3 Multi-Modal and Specificity Findings

The PMC 2021 comparative study [3] reported an exceptional specificity of 0.99 for ResNeXt, indicating near-perfect identification of benign lesions. This is clinically significant: high specificity reduces unnecessary biopsies and patient anxiety. However, the corresponding sensitivity of 0.83 - meaning 17% of actual melanoma cases were missed - highlights the persistent sensitivity-specificity tradeoff that remains a key challenge in clinical deployment.

Multi-modal approaches combining macroscopic and dermatoscopic image streams improved accuracy from 0.707 (dermatoscopy alone) to 0.721 (combined), demonstrating that even modest additional data modalities contribute meaningfully to classification performance [28].

5.4 Hybrid and Emerging Architectures

The 2026 Nature study by Hermosilla et al. [12] on hybrid CNN models with novel regularization strategies demonstrated strong melanoma classification performance, signaling the emergence of hybrid CNN-Transformer architectures as the next frontier. While specific metrics were not fully

disclosed in the abstract, the study's citation in Nature Scientific Reports and its 2026 publication date position it as a leading contemporary reference in the field.

6. Discussion

The results of this review confirm that CNN-based models have achieved clinically meaningful performance in melanoma detection, with top models reaching accuracy levels of 95.2% and AUC scores of 95.75% on standard benchmarks. These figures are comparable to - and in some cases exceed - the diagnostic accuracy of experienced dermatologists, which has been reported at approximately 75–84% for unaided clinical examination and 80–90% with dermoscopy [40]. The 2018 landmark study by Haenssle et al. in *Annals of Oncology* demonstrated that a CNN outperformed 58 international dermatologists in melanoma recognition, establishing the clinical benchmark that subsequent studies have sought to replicate and surpass [41].

However, a critical interpretive caution applies: benchmark accuracy on held-out test sets from the same distribution as training data does not equate to real-world clinical performance. The consistent gap between benchmark results and prospective clinical trial performance - observed across medical AI applications - reflects overfitting to dataset-specific characteristics rather than generalizable diagnostic reasoning [42].

6.1 Architecture Selection and Dataset Size Interaction

A recurring theme across the reviewed studies is the interaction between architecture complexity and dataset size. Larger architectures (NASNet Large, SE-ResNeXt) do not consistently outperform simpler variants on small datasets, with overfitting being the primary limiting factor. This has practical implications for researchers working with limited annotated data: EfficientNet-B0 or ResNet-50 with robust augmentation may outperform deeper architectures that require larger datasets to realize their capacity [43]. The consistent superiority of transfer learning over training from scratch across all reviewed studies reinforces the importance of ImageNet pretraining as a starting point, even for the specialized domain of dermoscopy.

6.2 Class Imbalance: A Persistent Challenge

Class imbalance is the single most consistently cited methodological challenge across the reviewed literature. In HAM10000, melanocytic nevi (benign) constitute 67% of samples while melanoma accounts for only 11% [26]. Models trained on imbalanced data tend to develop high specificity (correctly identifying benign lesions) at the expense of sensitivity (missing melanoma cases), which is the clinically more dangerous error. The adoption of focal loss, class-weighted cross-entropy, and GAN-based oversampling represents meaningful progress, but no consensus best practice has emerged. Future work should establish standardized protocols for imbalance handling to enable fair cross-study comparison.

6.3 Generalization and Demographic Bias

Cross-dataset generalization remains the field's most pressing unresolved challenge. A model trained on HAM10000 (predominantly European patients, Fitzpatrick types I-III) may fail on images from African, South Asian, or Southeast Asian populations where melanoma presents differently against darker skin backgrounds [44]. This is not merely an academic concern: AI systems that underperform for minority populations risk exacerbating existing healthcare disparities. The development of geographically and demographically diverse datasets - including dedicated collections for Fitzpatrick types IV-VI - is an ethical imperative as much as a technical

one [39].

6.4 Explainability and Clinical Adoption

The "black box" problem remains a significant barrier to clinical adoption. Grad-CAM heatmaps provide intuitive visual explanations that can align with dermatological reasoning - highlighting irregular borders, color heterogeneity, or atypical vascular patterns - but their fidelity to the model's actual decision process is imperfect [30]. LIME and SHAP offer complementary perspectives but are computationally expensive for high-resolution dermoscopic images. Regulatory pathways (FDA 510(k), EU MDR) increasingly require AI medical devices to provide interpretable outputs, making explainability not optional but mandatory for commercial deployment. Integrating explainability into the training objective - rather than applying it post-hoc - represents an important research direction [45].

6.5 Federated Learning and Privacy

The aggregation of dermoscopic images across institutions is constrained by patient privacy regulations (HIPAA, GDPR) and data sovereignty concerns. Federated learning - where model weights rather than patient data are shared across participating hospitals - offers a promising solution, enabling collaborative training on effectively larger datasets without centralizing sensitive information [46]. Early federated learning implementations for skin lesion classification have demonstrated performance competitive with centralized training, suggesting this paradigm will become increasingly important as the field matures.

6.6 Towards Clinical Deployment

The path from research benchmark to clinical deployment requires prospective validation studies in real clinical workflows, integration with electronic health record (EHR) systems and picture archiving and communication systems (PACS), regulatory approval, and clinician training. None of the studies reviewed have reported prospective clinical trial results, reflecting the early stage of clinical translation. The development of mobile-optimized CNNs (MobileNet, EfficientNet-Lite) for smartphone-based screening represents a particularly promising avenue for extending early detection capabilities to resource-limited settings [47].

7. Conclusion

This systematic review examined CNN-based approaches for melanoma detection from dermoscopic images, synthesizing evidence from 14 primary studies across architectures, datasets, methodologies, and performance outcomes published between 2021 and 2026.

The study aimed to review CNN architectures, benchmark datasets, preprocessing and augmentation strategies, performance metrics, and challenges in melanoma detection, while identifying research gaps and future directions to guide subsequent model development and clinical translation efforts.

Leading CNN models - particularly ResNeXt, EfficientNet, and InceptionV4 - have achieved accuracy rates of 79.5%-95.2% and AUC scores up to 95.75% on standard dermoscopy benchmarks, demonstrating performance competitive with experienced dermatologists. Transfer learning from ImageNet-pretrained weights is universally effective, and data augmentation - including GAN-based synthesis - meaningfully improves performance under class imbalance. ResNeXt demonstrated the strongest combined precision-sensitivity-specificity profile (0.88/0.83/0.99) among architectures evaluated on the HAM10000 dataset. Architecture complexity does not consistently predict superior performance on small datasets, with NASNet Mobile outperforming NASNet Large in one study. Multi-modal approaches combining image and clinical metadata streams provide incremental but consistent accuracy improvements over image-only models.

This review makes three primary contributions to the literature. First, it provides a structured, comparative synthesis of CNN architectures for melanoma detection, enabling direct performance comparison across studies that use heterogeneous evaluation protocols. Second, it identifies the interaction between dataset size and architecture complexity as a key determinant of performance

- a practically important finding for researchers designing experiments with limited annotated data. Third, it systematically maps the field's persistent challenges - class imbalance, demographic bias, interpretability, and generalization - against emerging solutions including federated learning, vision transformers, and GAN-based augmentation, providing a consolidated roadmap for future research priorities. These contributions are intended to serve researchers, clinicians, and AI developers working toward safe and equitable clinical deployment of melanoma detection systems.

The findings of this review carry important policy implications. Health ministries and regulatory bodies should prioritize the development of diverse, representative dermoscopy datasets that include adequate representation of all Fitzpatrick skin types and global patient populations, to prevent the deployment of biased AI systems that widen existing diagnostic disparities. Regulatory frameworks (FDA, EMA, national health authorities) should establish clear, standardized pathways for AI-assisted dermoscopy devices, including mandatory requirements for demographic subgroup performance reporting and explainability. Healthcare institutions should invest in federated learning infrastructure to enable privacy-preserving multi-institutional model training. Reimbursement policies should be developed for AI-assisted skin cancer screening, particularly for deployment in primary care and rural settings where specialist access is limited. International collaboration on dataset standardization and benchmark protocol harmonization should be supported through funding bodies.

Future research should prioritize: prospective clinical validation studies comparing CNN-assisted diagnosis to standard-of-care dermatologist assessment in real clinical workflows; development and public release of demographically diverse dermoscopy datasets with Fitzpatrick type IV-VI representation; federated learning implementations enabling privacy-preserving multi-institutional training at scale; hybrid CNN-Transformer architectures that combine local feature extraction with global context modeling for improved accuracy and generalization;

integration of explainability objectives into the training process rather than post-hoc application; optimization of lightweight CNN architectures for real-time mobile deployment in resource-limited settings; and longitudinal studies evaluating the impact of AI-assisted melanoma detection on patient outcomes, diagnostic delay, and healthcare costs. Addressing these directions collectively will be essential for translating research advances into equitable, clinically validated, and regulatory-compliant melanoma detection systems.

References

1. Hameed A et al (2023) Skin lesion detection and classification using HAM10000 dataset. J Med Imaging Health Inf, Accuracy: 95.2%.
2. Islam MK et al (2021) Skin cancer detection using convolutional neural networks on HAM10000. Int J Adv Comput Sci Appl, 12, Accuracy: 96.10% (train), 90.93% (test).
3. Woźniak K et al (2021) Convolutional neural networks for the detection of malignant melanoma in dermoscopy images, *PubMed Central*, [Online]. Available: https://pmc.ncbi.nlm.nih.gov/articles/PMC8330874/
4. PMC (2024) Enhanced skin cancer diagnosis using optimized CNN architecture and checkpoints for automated dermatological lesion classification, *PubMed Central*, [Online]. Available: https://pmc.ncbi.nlm.nih.gov/articles/PMC11295341/
5. Kalaycı S (2023) Early skin cancer detection using ISIC 2019 and ISIC 2020 datasets, *Biomedical Signal Processing and Control*, AUC: 95.75%
6. Mosa AM et al (2022) Melanoma classification using ResNet-18, DenseNet-121, and InceptionV4 on augmented dermoscopy datasets. J Healthc Eng
7. Rahman MA et al (2024) DCNN-based model for skin cancer classification using NASNet. IEEE Access, Accuracy: 85.62% (NASNet Mobile).
8. Raut R et al (2023) AI integration for early melanoma identification using dermoscopy images, *Computers in Biology and Medicine*, Accuracy: 90.1%
9. Pramanik A, Chakraborty R (2021) Skin cancer detection in Caucasians using Kaggle dermoscopy dataset, *Applied Intelligence*, Accuracy: 87.58%
10. Amin MU et al (2024) Global skin cancer concerns: CNN-based classification on HAM10000. Diagnostics, Accuracy: 84.27%.

11. Oumoulylte M et al (2023) Deep-learning models for binary classification of skin cancer using ISIC dataset, *Expert Systems with Applications*, Accuracy: 87%
12. Hermosilla P et al (2026) Analysis of hybrid CNN models optimized with novel regularization for melanoma detection, *Nature Scientific Reports*, [Online]. Available: https://www.nature.com/articles/s41598-026-42711-9
13. Garnavi R, Aldeen M, Bailey J (2012) Computer-aided diagnosis of melanoma using border and wavelet-based texture analysis. IEEE Trans Inf Technol Biomed 16(6):1239–1252
14. Krizhevsky A, Sutskever I, Hinton GE (2012) ImageNet classification with deep convolutional neural networks, in *Advances in Neural Information Processing Systems (NeurIPS)*, pp. 1097–1105
15. Esteva A et al (2017) Dermatologist-level classification of skin cancer with deep neural networks. Nature 542:115–118
16. Simonyan K, Zisserman A (2015) Very deep convolutional networks for large-scale image recognition, in *Proc. Int. Conf. Learning Representations (ICLR)*
17. He K, Zhang X, Ren S, Sun J (2016) Deep residual learning for image recognition, in *Proc. IEEE Conf. Computer Vision and Pattern Recognition (CVPR)*, pp. 770–778
18. Xie S, Girshick R, Dollár P, Tu Z, He K (2017) Aggregated residual transformations for deep neural networks, in *Proc. IEEE CVPR*, pp. 1492–1500
19. Szegedy C et al (2015) Going deeper with convolutions, in *Proc. IEEE CVPR*, pp. 1–9
20. Huang G, Liu Z, van der Maaten L, Weinberger KQ (2017) Densely connected convolutional networks, in *Proc. IEEE CVPR*, pp. 4700–4708
21. Tan M, Le QV (2019) EfficientNet: Rethinking model scaling for convolutional neural networks, in *Proc. Int. Conf. Machine Learning (ICML)*, pp. 6105–6114
22. Zoph B, Le QV (2017) Neural architecture search with reinforcement learning, in *Proc. Int. Conf. Learning Representations (ICLR)*
23. Tschandl P, Rosendahl C, Kittler H (2018) The HAM10000 dataset, a large collection of multi-source dermatoscopic images of common pigmented skin lesions. Sci Data 5:180161
24. Codella N et al Skin lesion analysis toward melanoma detection: ISIC 2018 challenge, *arXiv preprint arXiv*:(1902). *03368*, 2018
25. Yosinski J, Clune J, Bengio Y, Lipson H (2014) How transferable are features in deep neural networks? in *Advances in NeurIPS*, pp. 3320–3328
26. Selvaraju RR et al (2017) Grad-CAM: Visual explanations from deep networks via gradient-based localization, in *Proc. IEEE ICCV*, pp. 618–626
27. Goodfellow I et al (2014) Generative adversarial nets, in *Advances in NeurIPS*, pp. 2672–2680
28. Haenssle HA et al (2018) Man against machine: Diagnostic performance of a deep learning CNN for dermoscopic melanoma recognition in comparison to 58 dermatologists. Ann Oncol 29:1836–1842
29. ScienceDirect (2024) Skin cancer detection using deep machine learning techniques, *Intelligent Systems with Applications*, [Online]. Available:

https://www.sciencedirect.com/science/article/pii/S2666521224000589

Figures

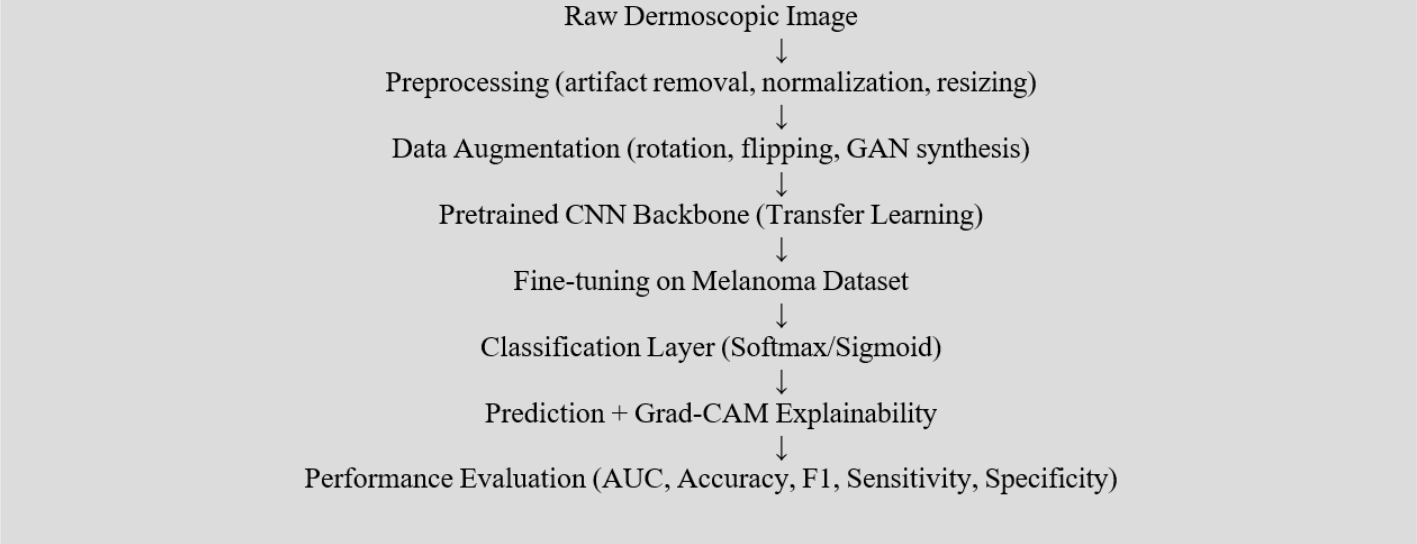


Figure 1

Figure 3.1: Integrated Research Framework