IJCRT.ORG ISSN : 2320-2882

INTERNATIONAL JOURNAL OF CREATIVE RESEARCH THOUGHTS (IJCRT)

An International Open Access, Peer-reviewed, Refereed Journal

Performance Evaluation of Edge-Enhanced CNNs and Pretrained Vision Transformer For Dermoscopic Skin Lesion Classification

Sandeep Kaur*, Er. Rupali

Department of Computer Science and Engineering, Guru Kashi University, Bathinda, Punjab, India

Abstract

Background: Skin cancer remains one of the most widespread malignancies globally, with over five million new diagnoses each year. Timely and precise identification plays a decisive role in improving patient survival. While conventional clinical assessment relies heavily on dermatologist expertise and is prone to inter-observer disagreement, artificial intelligence offers a promising avenue for standardized, automated detection. This study presents a dual-framework investigation into deep learning methods for binary dermoscopic image classification. The first framework evaluates a customconvolutional neural network alongside four edge-detection-enhanced variants that incorporate the Laplacian of Gaussian, Prewitt, Scharr, and Sobel operators. The second framework benchmarks five contemporary pretrained transformer-family architectures—Swin Transformer, ConvNeXt-Base, CAFormer-S18, MaxViT-Tiny, and CoaT-Lite-Medium—fine-tuned via transfer learning. Experiments were conducted on two publicly available dermoscopic datasets of differing scale: a curated HAM10000 subset (3,297 images) and the Kaggle Skin Cancer: Malignant vs. Benign collection (10,605 images). Among CNN variants, the Scharr operator consistently yielded the strongest results, achieving an AUC of 0.9564 on the larger dataset and the highest CNN accuracy of 79.55% on the smaller dataset. Among transformer-family models, ConvNeXt-Base attained the best overall accuracy of 90.61% with weighted precision, recall, and F1-score each reaching 0.90–0.91 on the HAM10000 subset.

Keywords: *Skin Cancer Detection; Vision Transformer; ConvNeXt; Edge Detection; Convolutional Neural Network; Transfer Learning; Dermoscopy; Medical Image Classification; ROC-AUC*

1. Introduction

The skin, being the body's largest organ, serves as the primary interface between internal physiology and the external environment. Structurally organized into the epidermis, dermis, and hypodermis, it performs vital functions, including physical protection, thermoregulation, sensory perception, and

vitamin D synthesis. Its persistent exposure to environmental stressors, particularly ultraviolet radiation, renders it vulnerable to a spectrum of disorders, of which skin cancer poses the most significant clinical threat.

Skin cancer manifests primarily in two broad categories: melanoma and non-melanoma. Melanoma, arising from melanocytes, is particularly aggressive and accounts for the majority of skin-cancer-related fatalities despite constituting a relatively small fraction of total diagnoses. Non-melanoma types, including basal cell carcinoma and squamous cell carcinoma, occur more frequently but are generally less lethal when identified early. With more than 5 million new cases reported worldwide each year, the imperative for early and reliable detection cannot be overstated. Conventional diagnostic workflows rely on clinical examination, dermoscopy, and histopathological biopsy. While dermoscopy enables visualization of subsurface morphological features otherwise invisible to the unaided eye, diagnostic accuracy remains heavily dependent on individual clinician expertise. Compounding this challenge, a well-documented shortage of trained dermatologists, especially in rural and resource-limited settings, creates significant access barriers to timely diagnosis.

Recent advances in deep learning have generated considerable momentum toward computer-aided diagnostic systems capable of analyzing dermoscopic images at high speed, with consistency and reproducibility. Several studies have demonstrated that carefully designed deep networks can achieve accuracy on par with, or even exceeding, that of board-certified dermatologists. However, persistent gaps remain in the literature: standard convolutional neural networks struggle to capture long-range spatial relationships; lesion boundary information, a critical indicator of malignancy, is rarely explicitly exploited; and systematic comparisons of modern transformer architectures under controlled conditions are scarce. This paper addresses these gaps through two complementary experimental frameworks. The first framework augments a custom CNN architecture with four classical edge detection operators, Laplacian of Gaussian (LoG), Prewitt, Scharr, and Sobel, and quantifies the effect of boundary-enhanced features on classification performance. The second framework evaluates five state-of-the-art pretrained transformer-family models under a standardized fine-tuning protocol. Both frameworks are validated across two dermoscopic datasets of different scales to assess generalisability. The principal contributions of this work include: (i) a controlled multi-operator edge detection comparison for CNN-based skin lesion classification, (ii) a rigorous five-model transformer benchmark for binary skin cancer detection, and (iii) cross-dataset evaluation providing generalisability insights absent from single-dataset studies.

2. Related Work

2.1 Classical Machine Learning Approaches

Early automated skin cancer detection methods employed traditional classifiers such as support vector machines, k-nearest neighbors, and random forests, operating on handcrafted features derived from color, texture, and shape descriptors. Ozkan et al. [1] compared several such classifiers and found that SVMs with radial basis function kernels outperformed alternatives among classical methods. Despite providing a foundation for automated analysis, these approaches were fundamentally constrained by their dependence on manually engineered feature representations, which could not capture the rich hierarchical information present in dermoscopic imagery.

2.2 Convolutional Neural Networks in Dermatology

The advent of deep convolutional neural networks marked a paradigm shift by enabling automatic hierarchical feature extraction directly from raw pixels. In a landmark study, Esteva et al. [2] trained a deep CNN on over 129,000 clinical images and achieved classification accuracy comparable to expert dermatologists for melanoma and keratinocyte carcinomas. Subsequent work explored architectures including VGGNet, ResNet, InceptionNet, DenseNet, and EfficientNet. Brinker et al. [3] showed that ResNet50 surpassed dermatologist performance in melanoma classification, while Ali et al. [4] demonstrated that EfficientNets achieve strong accuracy with fewer parameters. Hosseinzadeh et al. [5] proposed ensemble strategies that combine multiple deep models to improve robustness.

2.3 Transfer Learning for Medical Imaging

Transfer learning, adapting models pretrained on large-scale general datasets such as ImageNet to domain-specific medical imaging tasks, has become the dominant training paradigm in this field. This approach is especially valuable given the constraints on annotated clinical data arising from privacy regulations, high annotation costs, and the complexity of imaging. Haenssle et al.
[6] demonstrated that an InceptionV4-based system matched the diagnostic performance of 58 dermatologists, reinforcing the viability of pretrained networks for clinical decision support.

2.4 Vision Transformers and Modern Architectures

Dosovitskiy et al. [7] introduced the Vision Transformer (ViT), which processes images as sequences of patches via multi-head self-attention, enabling the capture of global dependencies that elude the local receptive fields of CNNs. Subsequent architectures refined this approach: the Swin Transformer [8] introduced hierarchical shifted-window attention for efficient multi-scale processing; ConvNeXt [9] modernized the convolutional paradigm through large-kernel depthwise convolutions and improved training strategies to rival transformer performance; CAFormer [10] hybridized pooling-based token mixing with channel attention; MaxViT [11] combined local window and global dilated attention; and CoaT [12] employed co-scale attention mechanisms. These architectures have shown competitive results on ImageNet and are beginning to be explored in medical image analysis, though systematic comparisons for skin cancer classification remain limited.

2.5 Edge Detection in Dermatological Image Analysis

Lesion boundary irregularity is among the most diagnostically significant features in dermoscopy, forming a core component of the ABCDE criteria (Asymmetry, Border, Color, Diameter, Evolving) used for malignancy assessment. Edge detection algorithms that emphasize boundary features have therefore been investigated as preprocessing enhancements. The Sobel, Prewitt, Scharr, and Laplacian of Gaussian operators each offer distinct tradeoff among noise sensitivity, edge localization accuracy, and computational cost. However, systematic comparison of multiple operators integrated into CNN classifiers under identical experimental conditions has not been adequately addressed in existing literature.

3. Materials and Methods

3.1 Datasets

Two publicly available dermoscopic datasets were used to evaluate all proposed frameworks, chosen to provide both a constrained-size and a larger-scale evaluation scenario.

Dataset-1 (HAM10000 Subset): A curated subset of the Human Against Machine with 10,000 training images, widely used as a benchmark in ISIC challenges. Images encompass diverse lesion types, including melanocytic nevi, melanoma, basal cell carcinoma, and others, reorganized into benign and malignant categories. The training partition contains 2,637 images, and the test partition contains 660 images, all at 224 × 224 pixel resolution.

Dataset-2 (Skin Cancer: Malignant vs. Benign): Sourced from Kaggle, this dataset comprises 9,605 training images (5,000 benign, 4,605 malignant) and 1,000 test images (500 per class), providing a balanced evaluation set. Original images at 300 × 300 pixels were resized to 224 × 224 for CNN experiments.

Table 1: Summary of Dataset Characteristics

Dataset	Train (Benign)	Train (Malignant)	Train Total	Test Total
Dataset-1 (HAM10000)	~1,350	~1,287	2,637	660
Dataset-2 (Malignant/Benign)	5,000	4,605	9,605	1,000

3.2 Data Preprocessing

All images were loaded in RGB format using the Python Imaging Library. For CNN models, pixel intensities were normalized to the [0, 1] range by dividing by 255.0, with no additional augmentation applied so as to establish conservative baselines. For transformer models, ImageNet-standard normalization (mean: 0.485, 0.456, 0.406; standard deviation: 0.229, 0.224, 0.225) was applied in accordance with pretraining conditions. Transformer training further employed an aggressive augmentation pipeline comprising random resized cropping (scale 0.6–1.0), horizontal and vertical flipping (probability 0.5 each), random rotation through 360 degrees, and color jitter (brightness, contrast, saturation: 0.3; hue: 0.05).

3.3 Edge Detection Operators

Four classical edge detection operators were integrated into the CNN framework to provide boundary-enhanced spatial representations as additional input features. Each operator applies a fixed, non-trainable convolutional kernel that highlights specific gradient characteristics of the input image.

- **Laplacian of Gaussian (LoG):** A second-order derivative method that first applies Gaussian smoothing to suppress noise, then computes the Laplacian to detect edges at zero-crossings. A

5×5 kernel was employed.

- **Prewitt Operator:** Computes horizontal and vertical gradient approximations using simplified 3×3 kernels with uniform weighting, providing computationally efficient first-order edge detection.
- **Scharr Operator:** An optimized variant of the Sobel operator that uses larger weight values (3, 10, 3 versus 1, 2, 1) to achieve improved rotational symmetry and more accurate gradient estimation.
- **Sobel Operator:** The most widely used gradient-based edge detector, computing horizontal and vertical image gradients through center-weighted 3×3 kernels.

3.4 CNN Architecture

The custom CNN follows a sequential design comprising two convolutional blocks. Each block contains two Conv2D layers (32 and 64 filters, 3×3 kernel, same padding, ReLU activation, Glorot uniform initialization) followed by 2×2 max pooling. After the convolutional blocks, the feature maps are flattened and passed through two dropout layers (rates 0.25 and 0.5) and two fully connected layers (512 units and 2-unit softmax output), yielding approximately 26.86 million trainable parameters. For edge-enhanced variants, fixed operator kernels are initialized in early convolutional layers to provide boundary-enhanced representations that supplement the learned features.

3.5 Transformer-Family Models

Five pretrained transformer-family architectures were evaluated, each offering a distinct approach to visual feature extraction:

Swin Transformer (swin_base_patch4_window7_224): Employs hierarchical shifted-window multi-head self-attention across four stages with progressively reduced spatial resolution. Pretrained on ImageNet-22K and fine-tuned on ImageNet-1K.

ConvNeXt-Base (convnext_base.fb_in22k_ft_in1k): A modernized pure convolutional network that matches transformer-level performance through 7×7 depthwise convolutions, inverted bottleneck design, reduced activation functions, and layer normalization. Pretrained on ImageNet-22K.

CAFormer-S18 (caformer_s18.sail_in22k_ft_in1k): A hybrid architecture that applies pooling-based token mixing in lower stages and self-attention in upper stages, balancing local and global feature extraction.

MaxViT-Tiny (maxvit_tiny_tf_224.in1k): Introduces multi-axis attention that combines local-window attention with global-dilated attention, enabling both local and global context modeling without quadratic complexity.

CoaT-Lite-Medium (coat_lite_medium.in1k): Uses co-scale attention with serial and parallel mechanisms to enable multiscale feature interactions across streams at different resolutions.

3.6 Training Configuration

CNN models were trained with the Adam optimizer using a batch size of 64. The basic CNN was trained for 50 epochs at a learning rate of 1×10^{-5}, while the edge-enhanced variants were trained for 25 epochs at a learning rate of 1×10^{-3}. A ReduceLROnPlateau scheduler (patience 5, factor 0.5) was

applied with categorical cross-entropy loss. Transformer models were fine-tuned using AdamW with weight decay of 0.01, a batch size of 32, an initial learning rate of 3×10^{-5}, and CosineAnnealingWarmRestarts scheduling. Label smoothing of 0.1 was applied. A two-phase protocol was followed: all five models were first compared over 15 epochs, then the best-performing model was extended to 40 epochs with early stopping (patience 3).

3.7 Evaluation Metrics

The performance was assessed using accuracy, precision, recall (sensitivity), F1-score, area under the receiver operating characteristic curve (AUC-ROC), and confusion matrices. These metrics collectively provide a comprehensive view of classifier behavior, particularly important in clinical settings where the costs of false negatives and false positives are asymmetric.

4. Results

4.1 CNN Results on Dataset-1

The basic CNN, trained for 50 epochs on Dataset-1, achieved a test accuracy of 78.48% and an AUC of 0.8504, correctly classifying 289 benign and 230 malignant cases out of 660 test samples. Among the edge-enhanced variants, performance varied considerably. The LoG operator produced the weakest results (accuracy 73.18%, AUC 0.7917), exhibiting high benign specificity but substantially reduced malignant sensitivity—a pattern attributable to the diffuse edge responses characteristic of second-order derivative operators. The Prewitt operator achieved the best AUC among edge variants at 0.8093, with notably strong malignant recall (0.78), though at the cost of reduced benign specificity. The Scharr operator performed closest to the baseline, achieving the highest CNN accuracy of 79.55% and an AUC of 0.8472, suggesting its optimized rotational symmetry provides edge representations that complement the learned convolutional features without introducing destructive interference. The Sobel operator yielded moderate results (accuracy: 75.15%; AUC: 0.8092).

Table 2: CNN Model Performance on Dataset-1

Model	Accuracy (%)	Precision	Recall	F1-Score	AUC
Basic CNN	78.48	0.76	0.77	0.76	0.8504
CNN + LoG	73.18	0.75	0.64	0.73	0.7917
CNN + Prewitt	73.03	0.68	0.78	0.76	0.8093
CNN + Scharr	79.55	0.78	0.77	0.77	0.8472
CNN + Sobel	75.15	0.73	0.72	0.72	0.8092

4.2 CNN Results on Dataset-2

On the larger Dataset-2, all CNN models exhibited substantially improved performance, reflecting the benefit of a 3.6-fold increase in training data. The basic CNN achieved 90.0% accuracy with an AUC of 0.9435. The Scharr-enhanced variant again attained the highest CNN AUC of 0.9564 with 90.3% accuracy, confirming its consistent advantage across datasets. The Sobel variant demonstrated an interesting trade-off: it achieved the highest malignant recall (447 true positives out of 500) at the cost of lower benign specificity (74 false positives), resulting in an overall accuracy of 87.3% despite a strong AUC of 0.9528.

Table 3: CNN Model Performance on Dataset-2

Model	TP	TN	FP	FN	Accuracy (%)	AUC
Basic CNN	437	463	37	63	90.0	0.9435
CNN + Scharr	437	466	34	63	90.3	0.9564
CNN + Sobel	447	426	74	53	87.3	0.9528

4.3 Transformer Model Comparison on Dataset-1

The five transformer-family models were evaluated on Dataset-1 through a 15-epoch comparison phase. ConvNeXt-Base and MaxViT-Tiny both achieved the highest validation accuracy of 91.06%, followed closely by CAFormer-S18 at 90.91%, CoaT-Lite-Medium at 90.30%, and Swin Transformer at 89.85%. ConvNeXt-Base was selected for extended training based on its lower validation loss at the tie-breaking point.

Table 4: Transformer Model Comparison (15-Epoch Phase, Dataset-1)

Model	Best Val. Accuracy (%)	Notes
Swin-Base	89.85	Peaked at epoch 6
ConvNeXt-Base	91.06	Best: early stop at epoch 12
CAFormer-S18	90.91	Steady improvement
MaxViT-Tiny	91.06	Fast convergence
CoaT-Lite-Medium	90.30	Peaked at epoch 9

Upon extended training to 40 epochs, ConvNeXt-Base reached its peak validation accuracy of 90.61% at epoch 7 before early stopping triggered at epoch 13. The final classification report reveals precision values of 0.88 and 0.94 for benign and malignant classes, respectively, recall of 0.95 and 0.85, and F1-scores of 0.92 and 0.89. The weighted average F1-score of 0.90 represents a substantial improvement over every CNN variant evaluated on the same dataset.

Table 5: ConvNeXt-Base Final Classification Report (Dataset-1)

Class	Precision	Recall	F1-Score	Support
Benign	0.88	0.95	0.92	360
Malignant	0.94	0.85	0.89	300
Weighted Avg.	0.91	0.90	0.90	660

4.4 Comprehensive Comparison

Table 6 consolidates results across all models and both datasets. The data reveal three key patterns: (a) edge detection operators modulate CNN performance in a dataset-dependent manner, with Scharr providing the most consistent improvement; (b) training data volume has a pronounced effect on CNN accuracy, with all CNN models benefiting substantially from the larger Dataset-2; and (c) transformer-based transfer learning on the smaller Dataset-1 achieves results comparable to CNN models trained on the much larger Dataset-2, demonstrating the capacity of pretrained representations to compensate for limited annotated medical data.

Table 6: Comprehensive Performance Summary Across All Frameworks

Model	Dataset	Accuracy (%)	AUC	F1 (Wt.)
Basic CNN	D-1	78.48	0.8504	~0.76
CNN + LoG	D-1	73.18	0.7917	~0.73
CNN + Prewitt	D-1	73.03	0.8093	~0.74
CNN + Scharr	D-1	79.55	0.8472	~0.77
CNN + Sobel	D-1	75.15	0.8092	~0.72
Basic CNN	D-2	90.00	0.9435	~0.90
CNN + Scharr	D-2	90.30	0.9564	~0.90
CNN + Sobel	D-2	87.30	0.9528	~0.87

5. Discussion

The experimental findings from this study illuminate several noteworthy aspects of deep learning-based skin lesion classification, with both scientific and practical significance.

First, the influence of edge detection operators on CNN performance is neither uniform nor predictable from theoretical properties alone. While the Scharr operator consistently delivered the strongest results across both datasets, the LoG operator, despite being the most mathematically sophisticated of the four, consistently underperformed. This can be attributed to LoG's tendency to produce diffuse, spatially spread edge responses that may obscure fine-grained texture information crucial for distinguishing benign from malignant lesions. By contrast, Scharr's optimized gradient kernels offer a better balance between edge localization precision and rotational invariance, making them more compatible with the downstream convolutional feature extraction pipeline. The Prewitt operator's notably high malignant

recall suggests it may be preferable in screening scenarios where missing a malignancy carries a greater clinical cost than generating a false alarm.

Second, the dramatic improvement in CNN performance from Dataset-1 to Dataset-2 underscores a well-established but often underemphasized reality in medical image analysis: training data volume remains a critical bottleneck for custom architectures trained from scratch. The basic CNN's accuracy jumped from 78.48% to 90.00% with a 3.6-fold increase in the training sample size, indicating that the network's capacity was not fully utilized on the smaller dataset. This observation carries direct practical implications for clinical deployment, where annotated dermoscopic datasets are often limited by privacy regulations, annotation costs, and institutional constraints.

Third, and perhaps most compellingly, transformer-based transfer learning largely circumvents the data scarcity problem. ConvNeXt-Base achieved 90.61% accuracy on Dataset-1 (2,637 training images), a performance comparable to that of CNN models with only 9,605 training images on Dataset-2. This demonstrates that rich representations learned during large-scale ImageNet pretraining transfer effectively to the dermoscopic domain, compensating for limited domain-specific annotations. The finding has immediate relevance for clinical settings in resource-limited environments where curating large labeled datasets is impractical.

Fourth, the observation that ConvNeXt-Base, a purely convolutional architecture, emerged as the top performer among the five transformer-family models is instructive. It suggests that the critical performance drivers are not self-attention mechanisms per se, but rather the combination of large-scale pretraining, modernized architectural design choices (large kernels, layer normalization, inverted bottleneck blocks), and effective fine-tuning strategies. This aligns with broader findings in the computer vision community, where ConvNeXt has repeatedly demonstrated transformer-competitive performance while retaining the computational simplicity of convolutions.

Fifth, from a clinical deployment perspective, ConvNeXt-Base's high benign recall of 0.95 is particularly valuable for screening applications, as it minimizes the risk of missed malignancies. Simultaneously, its malignant precision of 0.94 helps control the false positive rate, which directly affects patient anxiety and the burden of unnecessary biopsies. MaxViT-Tiny presents an interesting alternative for resource-constrained settings, matching ConvNeXt-Base's 15-epoch accuracy with fewer parameters, offering a potentially favorable accuracy-to-efficiency ratio.

When contextualized against prior work, the results are encouraging. Esteva et al. [2] reported dermatologist-level accuracy (~91%) using InceptionV3 on a private dataset of 129,000 images—a training set roughly 50 times larger than Dataset-1. The fact that ConvNeXt-Base achieves 90.61% on just 2,637 training images speaks to the considerable maturation of both model architectures and transfer learning methodologies over the intervening years.

6. Conclusion

This paper presents a systematic investigation of deep learning approaches for binary dermoscopic skin lesion classification, encompassing edge-detection-enhanced CNN architectures and five contemporary transformer-family models, evaluated across two datasets of differing scale.

The key findings are threefold. First, among edge detection operators, the Scharr operator delivers the

most consistent improvements over baseline CNN performance, achieving the highest CNN AUC of 0.9564 on the larger dataset. Second, transformer-based transfer learning decisively outperforms custom CNN baselines, with ConvNeXt-Base achieving 90.61% accuracy and a weighted F1-score of 0.90 on the HAM10000 subset—comparable to CNN performance on a dataset nearly four times larger. Third, cross-dataset evaluation confirms that the volume of training data is a critical bottleneck for custom CNNs, while pretrained architectures effectively compensate for limited annotated data.

These findings provide actionable guidance for the design of computer-aided dermoscopic diagnostic systems. For scenarios with constrained computational resources, Scharr-enhanced CNNs offer a lightweight option with measurable benefits for boundary features. Where accuracy is paramount, and adequate computational infrastructure is available, ConvNeXt-Base with transfer learning represents the strongest configuration evaluated in this study.

6.1 Limitations and Future Work

Several limitations should be acknowledged. The study addresses only binary classification; extension to multi-class subtype differentiation (e.g., melanoma vs. basal cell carcinoma) is warranted. Clinical metadata, such as patient demographics and lesion location, were not incorporated. An interpretability analysis using methods such as Grad-CAM or attention visualization was not performed an important consideration for clinical adoption. Furthermore, the transformer comparison was conducted only on Dataset-1; cross-dataset evaluation of transformer models remains an avenue for future work.

Promising future directions include multimodal fusion of image and clinical metadata, knowledge distillation for mobile deployment, federated learning for privacy-preserving multi-institutional training, integration with lesion segmentation pipelines, and prospective clinical validation against practicing dermatologists.

References

[1] I.A. Ozkan, M. Koklu, and I.U. Sert, “Skin lesion classification using machine learning algorithms,” in Proc. IDAP, 2017.

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

[3] T.J. Brinker, A. Hekler, A.H. Enk, et al., “Deep neural networks are superior to dermatologists in melanoma image classification,” European Journal of Cancer, vol. 119, pp. 11–17, 2019.

[4] K. Ali, Z.A. Shaikh, A.A. Khan, and A.A. Laghari, “Multiclass skin cancer classification using EfficientNets—a first step towards preventing skin cancer,” Neuroscience Informatics, vol. 2, no. 4, 100034, 2022.

[5] M. Hosseinzadeh, D. Hussain, F.M.Z. Mahmood, et al., “A model for skin cancer using combination of ensemble learning and deep learning,” PLOS ONE, vol. 19, no. 5, e0301275, 2024.

[6] H.A. Haenssle, C. Fink, R. Schneiderbauer, et al., “Man against machine: diagnostic performance of a deep learning CNN for dermoscopic melanoma recognition in comparison to 58 dermatologists,” Annals of Oncology, vol. 29, no. 8, pp. 1836–1842, 2018.

[7] A. Dosovitskiy, L. Beyer, A. Kolesnikov, et al., "An image is worth 16×16 words: transformers for image recognition at scale," in Proc. ICLR, 2021.

[8] Z. Liu, Y. Lin, Y. Cao, et al., "Swin Transformer: hierarchical vision transformer using shifted windows," in Proc. ICCV, 2021.

[9] Z. Liu, H. Mao, C.Y. Wu, et al., "A ConvNet for the 2020s," in Proc. CVPR, 2022.

[10] W. Yu, C. Si, P. Zhou, et al., "MetaFormer baselines for vision," arXiv preprint arXiv:2210.13452, 2022.

[11] Z. Tu, H. Talebi, H. Zhang, et al., "MaxViT: multi-axis vision transformer," in Proc. ECCV, 2022.

[12] W. Xu, Y. Xu, T. Chang, and Z. Tu, "Co-scale conv-attentional image transformers," in Proc. ICCV, 2021.

[13] M. Dildar, S. Akram, M. Irfan, et al., "Skin cancer detection: a review using deep learning techniques," Int. J. Environ. Res. Public Health, vol. 18, no. 10, 5479, 2021.

[14] D. Bisla, A. Choromanska, J.A. Stein, D. Polsky, and R. Berman, "Towards automated melanoma detection with deep learning: data purification and augmentation," arXiv preprint arXiv:1902.06061, 2019.

[15] M. Tan and Q. Le, "EfficientNet: rethinking model scaling for convolutional neural networks," in Proc. ICML, 2019.

[16] P. Tschandl, C. Rinner, Z. Apalla, et al., "Human-computer collaboration for skin cancer recognition," Nature Medicine, vol. 26, no. 8, pp. 1229–1234, 2020.

[17] Y. Liu, A. Jain, C. Eng, et al., "A deep learning system for differential diagnosis of skin diseases," Nature Medicine, vol. 26, no. 6, pp. 900–908, 2020.

[18] I. Krakowski, J. Kim, Z.R. Cai, et al., "Human-AI interaction in skin cancer diagnosis: a systematic review and meta-analysis," npj Digital Medicine, vol. 7, no. 1, 78, 2024.

[19] M.A. Kassem, K.M. Hosny, R. Damasevicius, and M.M. Eltoukhy, "Machine learning and deep learning methods for skin lesion classification and diagnosis: a systematic review," Diagnostics, vol. 13, no. 7, 1314, 2023.

[20] S. Maqsood and R. Damasevicius, "Multiclass skin lesion localization and classification using deep learning based features fusion and selection framework for smart healthcare," Neural Networks, vol. 160, pp. 238–258, 2023.

[21] N. Gessert, M. Nielsen, M. Shaikh, R. Werner, and A. Schlaefer, "Skin lesion classification using ensembles of multi-resolution EfficientNets with metadata," MethodsX, vol. 7, 100864, 2020.

[22] A. Naeem, T. Anees, M. Naqvi, and W.K. Loh, "A comprehensive analysis of recent deep and federated-learning-based methodologies for brain tumor and skin cancer diagnosis," Mathematics, vol. 11, no. 18, 3934, 2023.

[23] R. Zhang, "Melanoma detection using convolutional neural network and transfer learning," in Proc. IEEE Int. Conf. Consumer Electronics and Computer Engineering (ICCECE), 2023, pp. 75–79.

[24] M. Aljohani, R. Mustafa, and M. Bsharat, "Vision transformers for skin cancer classification: a comparative study with convolutional neural networks," IEEE Access, vol. 11, pp. 119654–119668, 2023.

[25] J. Höhn, A. Hekler, E. Krieghoff-Henning, et al., "Integrating patient data into skin cancer classification using convolutional neural networks: systematic review," Journal of Medical Internet Research, vol. 23, no. 7, e20708, 2021.

[26] G.M.S. Himel, M.M. Islam, K.A. Al-Aff, S.I. Karim, and M.K.U. Sikder, "Skin cancer segmentation and classification using vision transformer for automatic analysis in dermoscopy-based noninvasive digital system," International Journal of Biomedical Imaging, vol. 2024, 3909409, 2024.

[27] B. Cassidy, C. Kendrick, A. Brodzicki, J. Jaworek-Korjakowska, and M.H. Yap, "Analysis of the ISIC image datasets: usage, benchmarks and recommendations," Medical Image Analysis, vol. 75, 102305, 2022.

[28] S.K. Datta, M.A. Shaikh, S.N. Srihari, and M. Gao, "Soft attention improves skin cancer classification performance," in Proc. MICCAI Workshop on Interpretability of Machine Intelligence in Medical Image Computing (iMIMIC), 2021, pp. 13–23.