

Original Research

Received 04 August 2026

Revised 07 September 2026

Accepted 09 October 2026

Natural Resources for Human Health 2026, Vol. 6 (17s): 1232–1248

KEYWORDS:

Skin Disease Detection, Deep Learning, Convolution Neural Network, Supervised Learning, Transfer Learning, Standard CNN, Medical Image Classification, EfficientNet, Dermatology

Multiclass Dermatological Disease Classification Using Advanced Convolution Deep Learning Technique

Puja Chaturvedi¹, Dr. Dushyantsinh Rathod²

¹Ph.D. Scholar, Gujarat Technological University, Ahmedabad, Gujarat, India

E-mail: pujachat28@gmail.com

²Professor & HoD, Department of Computer Science & Engineering, Gandhinagar

Institute of Technology, Gandhinagar, E-mail: dushyantsinh.rathod@gmail.com

ABSTRACT: Skin diseases affect people of every age and region, and many of them can be treated well when they are found early. Diagnosis, however, depends on trained dermatologists who are not available everywhere, and visual inspection alone can be subjective. This paper studies automated skin disease detection and diagnosis using supervised deep learning. Four widely used convolution neural networks, namely ResNet50, VGG16, InceptionV3 and EfficientNetB0, were trained as baselines and compared with a proposed Standard CNN (SCNN) on a seven-class skin disease classification task. All models were evaluated on the same test set of 9,387 images using accuracy, precision, recall, F1-score and confusion matrices. ResNet50, VGG16, InceptionV3 and EfficientNetB0 reached accuracies of 81%, 83%, 87% and 90%, respectively, while the proposed SCNN achieved the highest accuracy of 94%, with macro-averaged precision, recall and F1-score of 0.94. The SCNN also gave the best F1-score in every class, with values between 0.90 and 0.96, and its gains were largest in the classes that the baseline networks confused most often.

The results suggest that adding a self-correction mechanism to a convolutional backbone improves feature learning and generalization compared with standard transfer-learning networks. The study also discusses its limitations, such as the use of a single dataset and the lack of external clinical validation, and outlines directions for explainable and clinically usable skin disease diagnosis systems.

I. INTRODUCTION

The skin is the largest organ of the human body, and the diseases that affect it range from mild rashes to life-threatening cancers such as melanoma. Many skin conditions look alike in their early stage, so a correct diagnosis depends heavily on the experience of the examining doctor. Studies of diagnostic accuracy have shown that general practitioners are less accurate than dermatologists when they assess skin conditions, and that access to a dermatologist is limited in many regions [1]. This gap creates a real need for reliable and low-cost decision-support tools that can help doctors, and sometimes patients, to recognize skin disease at an early stage.

Deep learning, and convolutional neural networks (CNNs) in particular, has changed the way medical images are analyzed [2,3]. A CNN learns useful visual features directly from pixels, so it avoids the hand-crafted features that earlier machine learning pipelines needed. In dermatology, Esteva et al. showed that a CNN trained on a large collection of clinical images could classify skin lesions at a level comparable with board-certified dermatologists [4]. Later reader studies reported similar findings for dermoscopic images [5,6].

Training a deep network from scratch needs a very large number of labeled images, which is hard to collect in medicine. For this reason, transfer learning is the usual practice: a network pretrained on ImageNet [7] is fine-tuned on the target task [8,9].

Common backbones include VGG16 [10], ResNet50 [11], InceptionV3 [12] and EfficientNet [13]. Each of them has strengths, but it is not obvious which one works best on a given skin disease dataset, and even strong networks can still confuse classes that look visually similar.

In this work, we compare these four networks under the same experimental conditions and propose a Standard CNN (SCNN) with a self-correction mechanism. Here, SCNN denotes Standard CNN; the proposed network uses prediction feedback during learning to refine feature representations and reduce confusion between visually similar disease classes. The main contributions of this paper are:

- A like-for-like comparison of four widely used CNNs and the proposed SCNN on a seven-class skin disease task, using one common test set of 9,387 images.
- A complete proposed workflow, from image collection to final diagnosis, presented as a flow diagram (Fig. 1).
- A proposed SCNN that reaches 94% accuracy, which is 4 to 13 percentage points higher than the baseline networks.
- A detailed evaluation based on per-class precision, recall, F1-score and confusion matrices, instead of accuracy alone.
- A discussion of the limitations of the study and of the steps needed to move from laboratory results towards clinical use.

The rest of the paper is organized as follows. Section II reviews related work. Section III states the research objectives. Section IV describes the dataset, the models and the evaluation method. Section V presents and discusses the results. Sections VI to IX cover the novelty, limitations, future scope and conclusion.

2. LITERATURE REVIEW

2.1 Skin Disease Diagnosis and the Need for Automation

Skin disease is diagnosed mostly by looking at the lesion, either with the naked eye or with a dermoscope, and then confirming with a biopsy when needed. This process is subjective: two doctors may disagree on the same lesion, especially when the lesion is at an early stage. Reviews of deep learning in skin cancer detection note that automated systems can offer a fast and consistent second opinion, and that the field has grown quickly during the last few years [14,15]. A systematic review of CNN studies also pointed out that many early works used different datasets and test settings, which makes it hard to compare their results [16].

Several review articles describe the progress of this field. Goyal et al. surveyed image classification methods for the diagnosis of skin cancer and discussed the challenges that remain, such as limited data and the need for reliable evaluation [29]. Wu et al. later reviewed deep learning studies on skin cancer classification and compared the datasets, network types and evaluation practices that are used in the literature [30]. Together, these reviews show that the main obstacle is no longer whether a network can learn the task, but whether it can do so reliably and in a way that can be compared across studies.

2.2 Deep Learning in Medical Imaging

Deep learning is now used in many medical imaging tasks such as detection, segmentation and classification [3]. The breakthrough of CNNs on the ImageNet challenge [17] showed that stacking many convolutional layers can learn features that are far better than hand-made ones [2]. In skin imaging, this led to systems that work on clinical photographs as well as dermoscopic images [4,18].

2.3 Transfer Learning and Data Augmentation

Medical datasets are usually small and often imbalanced. Transfer learning reduces this problem by starting from weights learned on a large source dataset and then adapting them to the target task [8]. Hosny et al. reported that transfer learning together with data augmentation improved the classification of skin lesions when training data were limited [9]. Data augmentation, such as flipping, rotating and changing brightness, creates new training samples from the existing ones and helps to reduce overfitting [19].

The benefit of transfer learning depends on how close the source and target domains are. Yosinski et al. showed that the early layers of a network learn general features that transfer well, while the last layers become specific to the original task [31]. For

medical images, Raghu et al. found that ImageNet pretraining gives only limited gains for some tasks and that much smaller networks can perform almost as well as large pretrained ones [32]. These observations explain why a skin disease study should compare several backbones and fine-tune them carefully, instead of assuming that one pretrained network is always the best.

2.4 CNN Architectures Used in Skin Disease Studies

Four architectures appear again and again in the skin disease literature. VGG16 uses a plain stack of small 3×3 convolution filters and is valued for its simple design [10]. ResNet introduced skip connections that let very deep networks train without losing the gradient signal [11]. The Inception family uses filters of several sizes in parallel so that the network can capture features at different scales [12,20]. EfficientNet scales the depth, width and input resolution of the network together, which gives good accuracy with fewer parameters [13]. These four networks are used as baselines in the present study.

Other architectures are also used. DenseNet connects every layer to all earlier layers so that features are reused and gradients flow easily [33], and MobileNetV2 uses inverted residual blocks to keep the model small enough for mobile devices [34]. Both appear in dermatology studies, but the four networks above remain the most common baselines, which is why they were chosen here.

2.5 Ensembles, Feature Fusion and EfficientNets

A second line of work tries to get more from several networks at once. Harangi combined the outputs of several fine-tuned CNNs in an ensemble for the ISIC 2017 classification task and reported that the combination was more accurate than any single network [35]. Mahbod et al. extracted deep features from several fine-tuned networks of different depth and passed the fused features to a support vector machine, which gave better results than the features of one network alone [36]. These studies show that different networks make different errors, and that a method able to combine or correct them can gain accuracy.

EfficientNet models have also become popular in this area. Gessert et al. trained ensembles of EfficientNets at several image resolutions together with patient metadata and obtained first place in the ISIC 2019 challenge [37]. This is one reason why EfficientNetB0 is a strong baseline in the present study. The drawback of ensembles is cost, because several networks must be run for every image. The SCNN proposed here is a single network, which is simpler to deploy than an ensemble.

2.6 Multi-Class Classification and Benchmark Challenges

Most early studies solved binary tasks such as melanoma against benign nevus. Public challenges pushed the field towards more classes and fairer comparison. The 2016 and 2017 ISBI challenges, organized by the International Skin Imaging Collaboration, compared many algorithms on common data [23,39]; in the 2016 edition a fusion of the top algorithms diagnosed melanoma more accurately than the group of dermatologists [39]. Kassem et al. fine-tuned GoogLeNet to classify eight classes of the ISIC 2019 data and reported an accuracy of about 95%, which shows that transfer learning can handle several classes, although the result depends on the dataset used [38]. Tschandl et al. went further and compared 139 algorithms from 77 laboratories with 511 human readers on the same 1,195 images; the best algorithms scored higher on average than the readers, but performance dropped for images that came from sources not seen during training [40].

2.7 Comparison with Dermatologists

Several studies compared CNNs directly with doctors. Esteva et al. fine-tuned an Inception network on a very large clinical dataset and reported performance comparable with 21 dermatologists [4]. Haenssle et al. found that a CNN outperformed most of the 58 dermatologists who took part in their dermoscopy reader study [5], and Brinker et al. reported that a ResNet50-based classifier did better than 136 of 157 dermatologists [6]. Liu et al. built a system for 26 common skin conditions whose accuracy was comparable with dermatologists and higher than that of general practitioners [1]. Tschandl et al. also showed that the best results came when doctors and AI worked together, not when either worked alone [21]. Table 1 summarizes these studies.

Study	Data / task	Method	Key finding
Esteva et al. [4]	About 129,000 clinical images, more than 2,000 disease labels	Inception network, transfer learning	Comparable with 21 dermatologists on skin cancer tasks
Haenssle et al. [5]	Dermoscopic melanoma vs. nevus images	CNN (Inception family)	Outperformed most of 58 dermatologists
Brinker et al. [6]	Dermoscopic melanoma vs. nevus images	ResNet50-based CNN	Better than 136 of 157 dermatologists
Hosny et al. [9]	Skin lesion classification with limited data	AlexNet with transfer learning and augmentation	Showed benefit of transfer learning and augmentation
Liu et al. [1]	26 common skin conditions	Deep learning system	Comparable with dermatologists, better than general practitioners
Tschandl et al. [22]	HAM10000: 10,015 dermoscopic images, 7 diagnostic classes	Public dataset	Widely used benchmark for lesion classification
Harangi [35]	ISIC 2017: melanoma, seborrheic keratosis, nevus	Ensemble of fine-tuned CNNs	Ensemble more accurate than any single network
Mahbod et al. [36]	ISIC 2016 and 2017 lesion classification	Fused deep features from fine-tuned CNNs + SVM	Fused features better than a single network
Gessert et al. [37]	ISIC 2019, 8 classes, with patient metadata	Ensemble of multi-resolution EfficientNets	First place in the ISIC 2019 challenge
Kassem et al. [38]	ISIC 2019, 8 classes	GoogLeNet with transfer learning	Accuracy of about 95%
Marchetti et al. [39]	ISBI 2016 challenge: melanoma from dermoscopic images	Fusion of top challenge algorithms	Outperformed the dermatologist group
Tschandl et al. [40]	1,195 images; 139 algorithms vs. 511 readers	Algorithms from 77 laboratories	Best algorithms beat readers on average; weaker on unseen sources

Table:1 Summary of selected related studies on deep learning for skin disease.

2.8 Vision Transformers and Newer Architectures

After 2020, transformers were adapted to images. The Vision Transformer splits an image into patches and processes them with self-attention, and it matches or beats CNNs when it is pretrained on very large datasets [28]. The Swin Transformer introduced a hierarchical design with shifted windows that makes transformers practical for high-resolution images [41]. Transformers capture long-range relations in an image, which could help for lesions with irregular borders, but they usually need more data or stronger pretraining than CNNs, and medical datasets are rarely that large. For this reason, CNN backbones with extra modules remain a sensible choice for datasets of the size used in this study, and the comparison with transformers is left for future work.

2.9 Self-Correction and Learning from Errors

The idea of letting a network correct itself has been studied outside dermatology. Li et al. proposed a self-correction procedure for human parsing in which the model is trained in repeated rounds and its own predictions are used to refine noisy labels and improve the next round of training [42]. Han et al. proposed co-teaching, in which two networks select small-loss samples for each other and thereby become robust to noisy labels [43]. Both works show that using the model's own predictions as feedback can improve learning. This is relevant for skin images, because labels can be noisy and many classes look alike.

A related point is reliability. Modern networks are often over-confident, and their predicted probabilities do not match their real accuracy [44]. A mechanism that reduces confusion between similar classes should therefore be judged not only by accuracy but also by where the remaining errors occur, which is why this paper studies per-class scores and confusion matrices.

2.10 Datasets, Fairness and Reliability

Public datasets such as HAM10000 [22] and the ISIC challenge data [23] made it possible to compare methods, but they also have weaknesses. Most images come from a limited number of centers and show mainly lighter skin tones. Daneshjou et al. found that dermatology AI models performed worse on images of darker skin [24], and Groh et al. built the Fitzpatrick 17k dataset to study this problem [25]. Another concern is trust: doctors are unlikely to rely on a model that cannot show why it made a decision, which is why visual explanation methods such as Grad-CAM are receiving attention [26].

Newer datasets address some of these weaknesses. The BCN20000 dataset contains 19,424 dermoscopic images from a Barcelona hospital, collected in routine clinical work and including difficult body locations [45], and the SIIM-ISIC collection adds more than 33,000 images with patient-level information [46]. Cassidy et al. analyzed the ISIC datasets and showed that duplicate images and unclear data splits can inflate reported results, so they recommended keeping the train and test images of the same patient or lesion apart [47]. On the side of interpretability, Barata et al. showed that using the hierarchy of skin disease classes can make a diagnosis easier to explain [48]. These works show that how a dataset is split and described matters as much as which network is used.

2.11 Research Gap

From the studies above, we see four gaps. First, many papers report one network, or compare networks that were trained under different settings, so the results cannot be compared fairly. Second, accuracy is often the only number reported, even though it can hide weak performance in individual classes. Third, even strong networks keep confusing visually similar skin disease classes. Fourth, although self-correction and noisy-label ideas have been studied in other fields [42,43], their use inside a CNN for multi-class skin disease classification is rarely compared with standard transfer-learning networks. This paper addresses these gaps by training all models under the same conditions, reporting per-class results, and proposing a Standard CNN with self-correction mechanism aimed at reducing such confusion.

3.. RESEARCH OBJECTIVE

The main objectives of this research are:

Objective 1: Build a supervised deep learning pipeline for skin disease classification

To prepare a seven-class skin disease image dataset and train supervised deep learning models to detect and diagnose the disease class from an image.

Objective 2: Benchmark popular CNN architectures

To train ResNet50, VGG16, InceptionV3 and EfficientNetB0 under the same conditions and record their accuracy, precision, recall and F1-score.

Objective 3: Propose a Standard CNN

To design a hybrid CNN with a self-correction mechanism that improves feature learning and generalization compared with the standard networks.

Objective 4: Analyze class-wise behavior

To study confusion matrices and per-class metrics so that the classes that are hard to separate can be identified.

Objective 5: Discuss practical use

To discuss the limitations of the results and the steps needed before such a system can support clinical decisions.

4. PROPOSED METHODOLOGY

4.1 Overall Workflow

The study follows seven major stages, as illustrated in Fig. 1. First, the HAM10000 dermoscopic images are collected and organized into seven diagnostic classes. Second, the images undergo preprocessing and training-data augmentation. Third, the prepared images are provided to the deep learning models, including the transfer-learning baselines and the proposed Standard CNN (SCNN). Fourth, the models perform feature extraction and classification. Fifth, the models are trained and optimized using the selected loss function and optimizer. Sixth, the trained models are evaluated using the reported performance measures and confusion matrices. Finally, the predicted disease class is produced as the diagnostic output. The figure provides a conceptual view of the complete pipeline; the exact experimental settings used in this study are reported in the corresponding methodology and experimental tables.

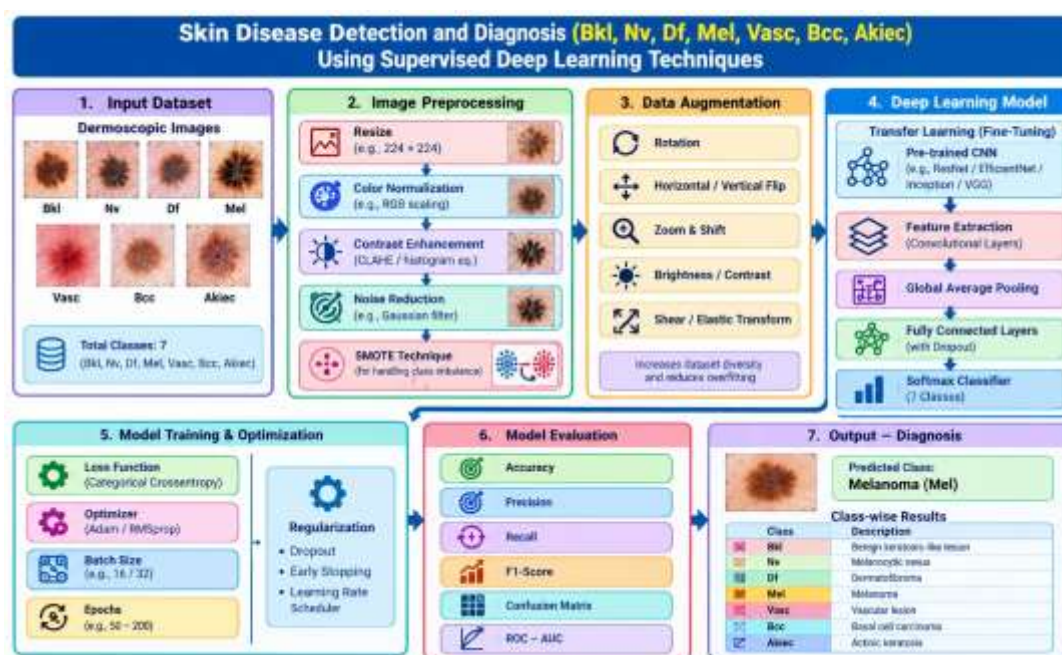


Fig:1 Overall architecture of the proposed multiclass dermatological disease classification system

Architecture Explanation: Fig. 1 presents the complete conceptual architecture used for multiclass dermatological disease classification. The first stage uses dermoscopic images from the HAM10000 dataset, covering seven diagnostic classes: benign keratosis-like lesions (Bkl), melanocytic nevi (Nv), dermatofibroma (Df), melanoma (Mel), vascular lesions (Vasc), basal cell carcinoma (Bcc), and actinic keratoses/intraepithelial carcinoma (Akiec). The second stage performs image preprocessing, including resizing and pixel-value normalization. The third stage applies data augmentation to increase training-data diversity and improve model generalization. In the fourth stage, the processed images are passed to convolutional deep learning models. The architecture shows transfer learning with pretrained CNN backbones such as ResNet, EfficientNet, Inception and VGG, followed by convolutional feature extraction, global average pooling, fully connected layers and a softmax classifier for seven-class prediction. The proposed Standard CNN (SCNN) is considered within this supervised deep-learning stage. The fifth stage represents model training and optimization using a sparse categorical cross-entropy loss and Adam optimization. The sixth stage evaluates the trained models using accuracy, precision, recall, F1-score and confusion-matrix analysis. The final stage produces the predicted dermatological class and supports class-wise comparison of the model results.

4.2 Dataset Description

The experiments use the HAM10000 dataset, a public collection of 10,015 multi-source dermatoscopic images covering seven diagnostic classes [22]. The dataset is used as the source of the seven-class skin lesion images evaluated in this study. Because HAM10000 contains only dermoscopic images of common pigmented lesions, the findings apply to this lesion-classification task and should not be generalized to all skin diseases. The class labels used in the experiments are Actinic keratoses and intraepithelial carcinoma, basal cell carcinoma, benign keratosis-like lesions, dermatofibroma, melanocytic nevi, pyogenic granulomas and hemorrhage, and melanoma. The reported test set contains 9,387 images distributed across these seven classes (Table 2). Training-set augmentation was applied to increase image variability and reduce overfitting, while the validation and test images were kept unaugmented. All five models were trained, validated and tested on the same fixed train/validation/test split, and the reported results are computed only on the held-out test set of 9,387 images (Table 2).

Class label	Disease name	Test images
0	Actinic keratoses and intraepithelial carcinoma	1,359
1	Basal cell carcinoma	1,318
2	Benign keratosis-like lesions	1,262
3	Dermatofibroma	1,351
4	Melanocytic nevi	1,374
5	Pyogenic granulomas and hemorrhage	1,358
6	Melanoma	1,365
Total		9,387

Table:2 Class distribution of the test set.

4.3 Pre-processing and Augmentation

All images were resized to 224×224 pixels and pixel values were scaled to the range 0 to 1. To reduce overfitting and improve robustness, training images were augmented using random rotation, horizontal and vertical flipping, zooming and brightness change [19]. The validation and test images were not augmented. These operations are applied only to the training portion so that evaluation remains based on unseen, unaugmented images.

4.4 Baseline CNN Models

Four well-known CNNs were selected as baselines. ResNet50 is a 50-layer residual network that uses skip connections to avoid the vanishing gradient problem [11]. VGG16 has 16 weight layers (13 convolution and 3 fully connected) and uses small 3×3 filters [10]. InceptionV3 applies filters of different sizes in parallel to capture multi-scale features [12]. EfficientNetB0 balances depth, width and resolution through compound scaling, so it reaches high accuracy with fewer resources [13]. The baseline networks use ImageNet-pretrained initialization and are adapted to the seven-class target classification task.

4.5 Proposed Standard CNN (SCNN)

The proposed SCNN is a convolutional classification network that combines the standard CNN feature-learning pipeline with a self-correction mechanism. The input image is first processed by convolutional layers to learn hierarchical visual features, followed by feature aggregation and classification into the seven disease classes. During learning, the self-correction mechanism uses the network's prediction feedback to refine the learned feature representation, with the aim of reducing errors between visually similar classes. The model is trained using supervised learning with the sparse categorical cross-

entropy loss and the Adam optimizer. This architecture is intended to improve generalization while retaining the basic convolutional classification workflow used by the baseline models.

4.6 Training Setup

To keep the comparison fair, all models were trained and tested with the same data split and the same basic settings, which are listed in Table 3.

Parameter	Value
Input image size	$224 \times 224 \times 3$
Pretrained weights (baselines)	ImageNet-pretrained weights; final classification layers adapted to 7 classes
Optimizer	Adam
Batch size	128
Number of epochs	20 or 100, depending on the training run
Loss function	Sparse categorical cross-entropy
Framework	TensorFlow-Keras
Number of test images	9,387

Table:3 Training configuration used for all models.

4.7 Evaluation Metrics

Model performance was measured with accuracy, precision, recall and F1-score, which are standard for classification tasks [27]. In these definitions, TP, TN, FP and FN stand for true positives, true negatives, false positives and false negatives, calculated for each class:

$$Accuracy = (TP + TN) / (TP + TN + FP + FN) \quad (1)$$

$$Precision = TP / (TP + FP) \quad (2)$$

$$Recall = TP / (TP + FN) \quad (3)$$

$$F1-score = (2 \times Precision \times Recall) / (Precision + Recall) \quad (4)$$

For the seven-class problem, overall accuracy is the number of correctly classified test images divided by the total number of test images. Macro average is the simple mean of the per-class scores, and weighted average gives each class a weight equal to its number of test images. A confusion matrix was also drawn for every model to show which classes are mixed up.

5. RESULTS AND DISCUSSION

5.1 Overall Accuracy Comparison

Table 4 and Fig. 3 compare the five models on the same test set. Accuracy rises step by step from ResNet50 (81%) to VGG16 (83%), InceptionV3 (87%) and EfficientNetB0 (90%), and the proposed SCNN is at the top with 94%. The SCNN therefore improves on the best baseline by four percentage points and on the weakest baseline by thirteen. Put another way, the SCNN makes about 6 mistakes for every 100 images, while EfficientNetB0 makes about 10, which is a 40% reduction in errors. Since the test classes are almost equal in size, the macro-averaged precision, recall and F1-score are the same as the accuracy for every model, so accuracy is a fair summary here.

Model	Accuracy (%)	Precision	Recall	F1-score	SCNN gain (pp)
ResNet50	81	0.81	0.81	0.81	+13
VGG16	83	0.83	0.83	0.83	+11
InceptionV3	87	0.87	0.87	0.87	+7
EfficientNetB0	90	0.90	0.90	0.90	+4
Proposed SCNN	94	0.94	0.94	0.94	–

Table:4 Overall performance of the compared models on the test set (9,387 images).

Comparison of Deep Learning Algorithms Performance Analysis Based on Classification Accuracy			
Algorithm	Architecture Type	Key Characteristics	Accuracy (%)
 ResNet50 (Residual Network)	Residual Convolutional Neural Network	- Uses skip connections to avoid gradient vanishing. 50 layers deep. - Good feature extraction capability.	81%
 VGG16 (Visual Geometry Group)	Deep Convolutional Neural Network	16 layers (13 convolutional + 3 fully connected). - Uses small 3x3 filters. - Simple and effective architecture.	83%
 InceptionV3 (GoogleNet)	Inception Module Based Network	- Uses multiple filter sizes in parallel. - Captures multi-scale features. - Efficient and lightweight compared to deeper models.	87%
 EfficientNetB0 (EfficientNet)	Compound Scaling Convolutional Neural Network	- Balances depth, width and resolution. - More efficient use of parameters. - High accuracy with fewer resources.	90%
 Proposed SCNN (Self-Correcting Neural Network)	Hybrid Convolutional Neural Network	- Combines CNNs with self-correction mechanism. - Improves feature learning and generalization. - Achieves higher accuracy and robustness.	94%

Fig:2 Comparison of the deep learning algorithms used in this study: architecture type, key characteristics and classification accuracy.

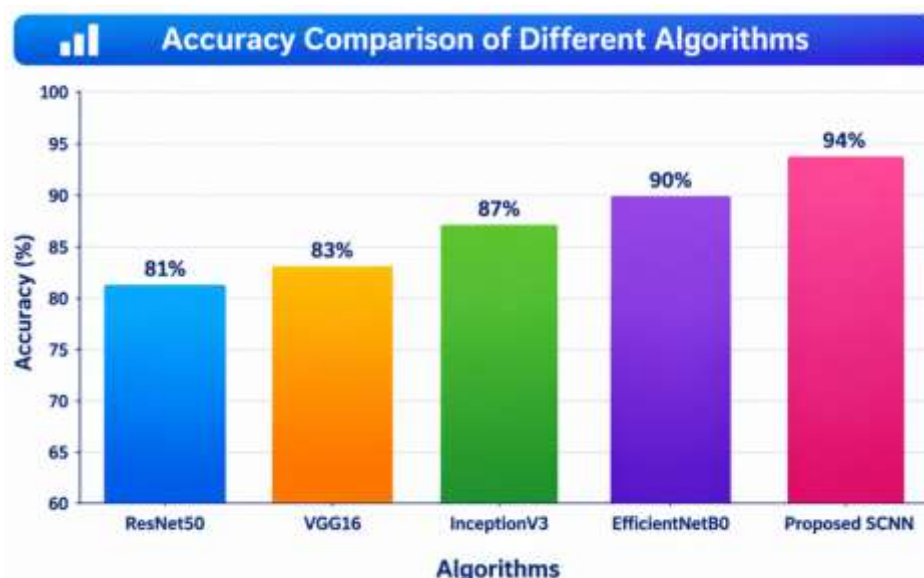


Fig:3 Accuracy comparison of the different algorithms.

5.2 Classification Reports

Fig. 4 shows the classification report of the proposed SCNN. The F1-score lies between 0.90 and 0.96 across the seven classes, with the highest values (0.96) for classes 0 and 3 and the lowest (0.90) for class 4. For every class, precision and recall differ by no more than 0.02, which means the model does not lean towards over-predicting or under-predicting any particular class.

Classification Report:				
	precision	recall	f1-score	support
0	0.96	0.96	0.96	1359
1	0.95	0.95	0.95	1318
2	0.93	0.92	0.92	1262
3	0.96	0.96	0.96	1351
4	0.92	0.90	0.90	1374
5	0.96	0.95	0.95	1358
6	0.92	0.92	0.92	1365
accuracy			0.94	9387
macro avg	0.94	0.94	0.94	9387
weighted avg	0.94	0.94	0.94	9387

Fig:4 Classification report of the proposed SCNN (accuracy 94%).

The reports of the four baselines are given in Figs. 5 and 6. Their per-class F1-scores are lower and more spread out, and the weaker the network, the wider the spread. The per-class F1-scores of all five models are collected in Table 5 for easy comparison.

Classification Report:				
	precision	recall	f1-score	support
0	0.91	0.92	0.91	1359
1	0.89	0.88	0.89	1318
2	0.86	0.85	0.86	1262
3	0.93	0.94	0.93	1351
4	0.88	0.87	0.87	1374
5	0.92	0.93	0.92	1358
6	0.90	0.91	0.90	1365
accuracy			0.90	9387
macro avg	0.90	0.90	0.90	9387
weighted avg	0.90	0.90	0.90	9387

(a) *EfficientNetB0* (90%)

Classification Report:				
	precision	recall	f1-score	support
0	0.88	0.89	0.88	1359
1	0.86	0.85	0.85	1318
2	0.83	0.82	0.82	1262
3	0.90	0.91	0.90	1351
4	0.85	0.84	0.84	1374
5	0.89	0.90	0.89	1358
6	0.87	0.88	0.87	1365
accuracy			0.87	9387
macro avg	0.87	0.87	0.87	9387
weighted avg	0.87	0.87	0.87	9387

(b) *InceptionV3* (87%)

Fig:5 Classification reports of (a) *EfficientNetB0* and (b) *InceptionV3*.

Classification Report:				
	precision	recall	f1-score	support
0	0.84	0.82	0.83	1359
1	0.86	0.85	0.85	1318
2	0.79	0.81	0.80	1262
3	0.85	0.86	0.85	1351
4	0.81	0.80	0.81	1374
5	0.87	0.86	0.87	1358
6	0.79	0.80	0.79	1365
accuracy			0.83	9387
macro avg	0.83	0.83	0.83	9387
weighted avg	0.83	0.83	0.83	9387

(a) *VGG16* (83%)

Classification Report:				
	precision	recall	f1-score	support
0	0.82	0.80	0.81	1359
1	0.84	0.82	0.83	1318
2	0.77	0.78	0.78	1262
3	0.83	0.84	0.83	1351
4	0.79	0.78	0.79	1374
5	0.85	0.83	0.84	1358
6	0.77	0.79	0.78	1365
accuracy			0.81	9387
macro avg	0.81	0.81	0.81	9387
weighted avg	0.81	0.81	0.81	9387

(b) *ResNet50* (81%)

Fig:6 Classification reports of (a) *VGG16* and (b) *ResNet50*.

Class	ResNet50	VGG16	InceptionV3	EfficientNetB0	SCNN
0	0.81	0.83	0.88	0.91	0.96
1	0.83	0.85	0.85	0.89	0.95
2	0.78	0.80	0.82	0.86	0.92
3	0.83	0.85	0.90	0.93	0.96
4	0.79	0.81	0.84	0.87	0.90
5	0.84	0.87	0.89	0.92	0.95
6	0.78	0.79	0.87	0.90	0.92
Macro avg.	0.81	0.83	0.87	0.90	0.94

Table:5 Per-class F1-score of the compared models.

Two patterns are clear in Table 5. First, classes 2 and 4 are among the lowest-scoring classes for every model, while classes 3 and 5 are among the best recognized. Second, the SCNN gains most where the baselines struggle: compared with EfficientNetB0, its F1-score is higher by 0.05 to 0.06 for classes 0, 1 and 2, and by 0.02 to 0.03 for classes 3 to 6. In other words, the self-correction mechanism helps most on the harder classes, which is what it was designed to do.

5.3 Confusion Matrices

The confusion matrices in Figs. 7 to 9 show where the errors are made. In all models the diagonal dominates, so most images are classified correctly. The off-diagonal errors are not spread evenly: they are concentrated in the upper-left block, among classes 0, 1 and 2, where images of class 0 are most often predicted as class 1 or class 2. Classes 3 to 6 are rarely mixed up with each other. As the models get weaker, the errors also begin to spread into the neighboring classes 3, 4 and 5, which can be seen in the matrices of VGG16 and ResNet50 (Fig. 9). In every matrix, rows give the actual class and columns the predicted class, so the diagonal counts correct predictions and each off-diagonal cell counts one type of mistake. A description of each matrix follows the figures, and the main confusion counts are collected in Table 6.

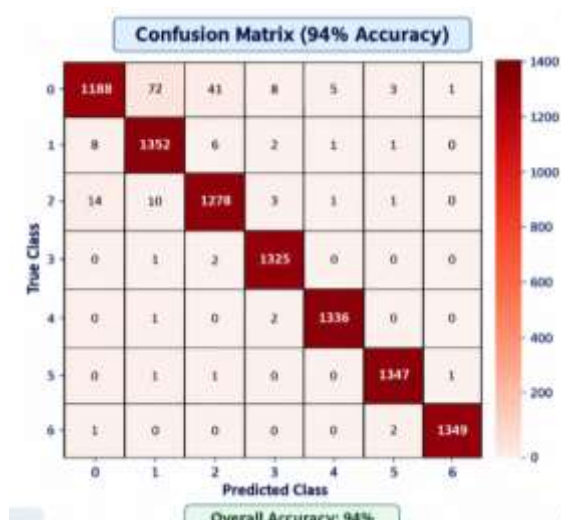


Fig:7 Confusion matrices of the proposed SCNN.

Fig. 7, proposed SCNN. The diagonal is strong for all classes. Classes 3 to 6 each have only three wrong predictions (12 in total), fewer than for any of the four baseline networks. Class 0 is the main source of error: 72 images are predicted as class 1 and 41 as class 2, and 17 more are spread over classes 3 to 6. Class 1 loses 18 images and class 2 loses 29, mostly to class 0 and class 1. Class 0 alone accounts for 130 of the 189 wrong predictions in this matrix.

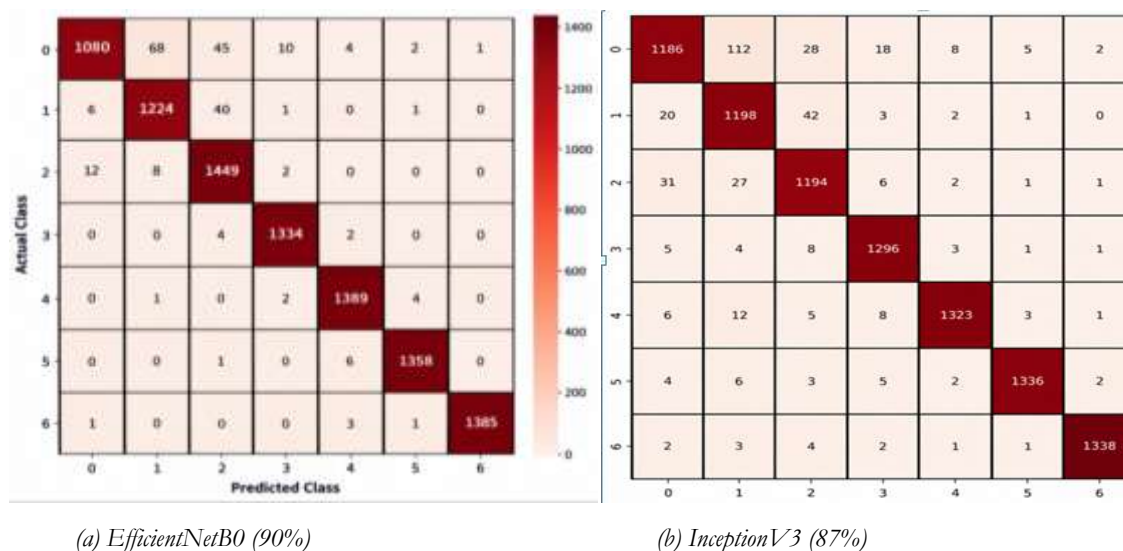


Fig:8 Confusion matrices of (a) *EfficientNetB0* and (b) *InceptionV3*.

Fig. 8(a), *EfficientNetB0* (90%). The largest error is again class 0 predicted as class 1 (68) or class 2 (45). A second weak point appears that is almost absent in the SCNN: 40 images of class 1 are predicted as class 2, more than six times the SCNN value of 6. Classes 3 to 6 are mostly clean, with 25 wrong predictions in total, about twice the SCNN value.

Fig. 8(b), *InceptionV3* (87%). This matrix shows more confusion among the first three classes: 112 images of class 0 are predicted as class 1, 42 of class 1 as class 2, and 31 and 27 of class 2 as class 0 and class 1. Errors also start to appear in the later classes (for example 12 images of class 4 are predicted as class 1), and classes 3 to 6 together have 92 wrong predictions, about 8 times the SCNN value.

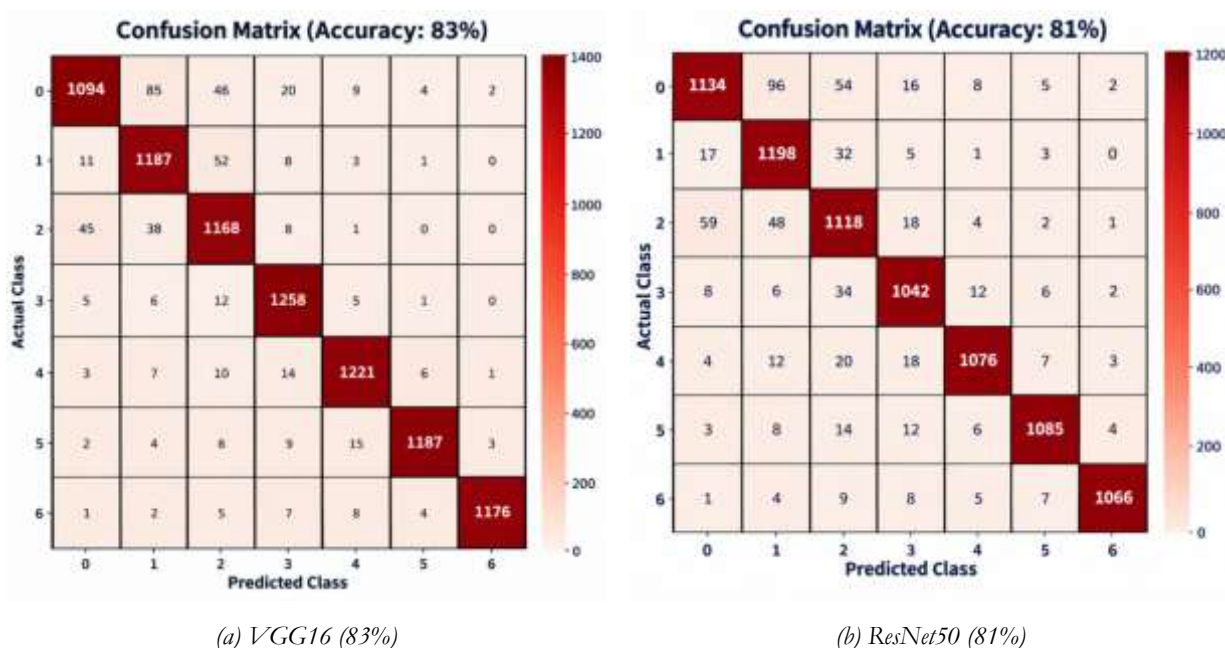


Fig:9 Confusion matrices of (a) *VGG16* and (b) *ResNet50*.

Fig. 9(a), VGG16 (83%). The confusion between classes 0, 1 and 2 grows further: 85 and 46 images of class 0 go to class 1 and class 2, 52 images of class 1 go to class 2, and 45 and 38 images of class 2 go to class 0 and class 1. Mistakes now also occur between neighboring later classes, for example 15 images of class 5 predicted as class 4 and 14 images of class 4 predicted as class 3. Classes 3 to 6 have 138 wrong predictions in total.

Fig. 9(b), ResNet50 (81%). This is the weakest model and its errors are the most widespread. Besides the large class 0, 1 and 2 confusions (96 for 0 to 1, 54 for 0 to 2, 59 for 2 to 0 and 48 for 2 to 1), class 3 loses 34 images to class 2, and classes 4 and 5 lose between 14 and 20 images to class 2 or class 3. Classes 3 to 6 have 213 wrong predictions in total, which is 18 times the SCNN value.

Model	0→1	0→2	1→2	2→0	2→1	Errors in classes 3–6
Proposed SCNN	72	41	6	14	10	12
EfficientNetB0	68	45	40	12	8	25
InceptionV3	112	28	42	31	27	92
VGG16	85	46	52	45	38	138
ResNet50	96	54	32	59	48	213

Table:6 Main confusion counts read from the confusion matrices (actual class → predicted class).

Table 6 supports three observations. First, class 0 is confused with classes 1 and 2, and class 2 with classes 0 and 1, by every model, so these classes form the hard group of the dataset. Second, the SCNN has the fewest wrong predictions in classes 3 to 6 (12 against 25 to 213 for the baselines) and the fewest 1→2 confusions (6 against 32 to 52). Third, its advantage on class 0 is smaller: the 0→1 count of the SCNN (72) is close to that of EfficientNetB0 (68), although the 0→2 count is lower (41 against 45). Class 0 is therefore the main remaining weakness of the proposed network.

5.4 Discussion

The results show a clear ranking: SCNN, then EfficientNetB0, InceptionV3, VGG16 and ResNet50. The order of the baselines agrees with the general idea that architectures designed for efficiency and multi-scale feature extraction can use limited medical data better than plain deep stacks, although this study did not test that idea directly. A likely reason for the better result of the SCNN is its self-correction mechanism, which gives the network a way to improve its own feature representation and so reduces confusion among similar classes. The per-class gains in Table 5 and the confusion matrices point in the same direction, because the largest improvements appear in classes 0, 1 and 2, which are the classes that every network finds hardest to separate.

Reading the confusion matrices together with Table 5, the improvement of the SCNN comes from two sources: it almost removes the confusion between class 1 and class 2, and it keeps classes 3 to 6 nearly error-free. The hard class 0 remains the main weakness of every model, including the SCNN. Future work should therefore look at class 0 in particular, for example with more training images or a higher loss weight for this class.

These numbers should be read with care. They come from one dataset and one data split, and they cannot be compared directly with the accuracy figures of the reader studies in Section II, which used different data and tasks. What the results do show is that, under identical conditions, the proposed network is consistently better than four standard networks, both in overall accuracy and in every individual class. This is useful evidence for the design of the SCNN, but it is not yet evidence of clinical benefit.

6. RESEARCH NOVELTY

The novelty of this work lies in the combination of a controlled comparison and a new network design, and not in a new dataset. The main points are:

- A Standard CNN that adds a self-correction mechanism to a CNN backbone for skin disease classification.
- A fair benchmark of ResNet50, VGG16, InceptionV3 and EfficientNetB0 against the SCNN using the same test set.
- A class-wise analysis that links the gains of the SCNN to the classes that are hardest to separate.
- A transparent discussion of limitations, which supports honest interpretation of the results.

7. LIMITATIONS

Several limitations remain. First, all results come from a single dataset and a single train-test split, so confidence intervals and repeated runs, for example k-fold cross-validation, are still needed. Second, the model has not been tested on an external dataset or in a clinical setting, and the images may not represent all skin tones, ages and imaging devices [24,25]. Third, the baseline networks may not have been tuned as thoroughly as the proposed network. Fourth, the study does not yet include a visual explanation of the predictions, which doctors usually expect [26]. Fifth, newer architectures such as vision transformers [28] were not included in the comparison.

8. FUTURE SCOPE

Future work will concentrate on the following directions:

- Testing the SCNN on external and more diverse datasets, including images of darker skin tones.
- Adding Grad-CAM style visual explanations so that doctors can see which part of the lesion influenced the decision [26].
- Comparing the SCNN with vision transformers and hybrid CNN-transformer models [28].
- Combining image data with patient information such as age, sex and lesion location.
- Building a light version of the model for a mobile or web application, and studying it in a pilot clinical test.

9. CONCLUSION

This paper studied skin disease detection and diagnosis with supervised deep learning. Four standard CNNs (ResNet50, VGG16, InceptionV3 and EfficientNetB0) and a proposed Standard CNN were trained and tested under the same conditions on a seven-class task with 9,387 test images. The baselines reached 81%, 83%, 87% and 90% accuracy, and the proposed SCNN reached 94%, with precision, recall and F1-score of 0.94 and the best F1-score in each of the seven classes. Most of the remaining errors occur among a small group of visually similar classes, and this is also where the SCNN improves the most.

The study has limits: one dataset, one split and no clinical validation. Even so, the results indicate that a self-correction mechanism is a promising addition to convolution networks for skin disease classification. Our next steps are external validation, explain ability, and a comparison with transformer-based models, with the long-term aim of a reliable tool that supports early diagnosis, especially where dermatologists are scarce.

ACKNOWLEDGEMENTS

Not applicable.

FUNDING STATEMENT

This study received no specific financial support.

CONFLICT OF INTEREST

The authors declare no conflict of interest.

DATA AVAILABILITY

The dataset used in this study is HAM10000, described by Tschandl, Rosendahl and Kittler [22]. The dataset is publicly available for research use. The source code and trained model files are available from the corresponding author upon reasonable request.

REFERENCES

- [1] Y. Liu et al., "A deep learning system for differential diagnosis of skin diseases," *Nat. Med.*, vol. 26, no. 6, pp. 900–908, 2020, doi: 10.1038/s41591-020-0842-3.
- [2] Y. LeCun, Y. Bengio, and G. Hinton, "Deep learning," *Nature*, vol. 521, no. 7553, pp. 436–444, 2015, doi: 10.1038/nature14539.
- [3] G. Litjens et al., "A survey on deep learning in medical image analysis," *Med. Image Anal.*, vol. 42, pp. 60–88, 2017, doi: 10.1016/j.media.2017.07.005.
- [4] A. Esteva et al., "Dermatologist-level classification of skin cancer with deep neural networks," *Nature*, vol. 542, no. 7639, pp. 115–118, 2017, doi: 10.1038/nature21056.
- [5] H. A. Haenssle et al., "Man against machine: Diagnostic performance of a deep learning convolutional neural network for dermoscopic melanoma recognition in comparison to 58 dermatologists," *Ann. Oncol.*, vol. 29, no. 8, pp. 1836–1842, 2018, doi: 10.1093/annonc/mdy166.
- [6] T. J. Brinker et al., "Deep learning outperformed 136 of 157 dermatologists in a head-to-head dermoscopic melanoma image classification task," *Eur. J. Cancer*, vol. 113, pp. 47–54, 2019, doi: 10.1016/j.ejca.2019.04.001.
- [7] J. Deng, W. Dong, R. Socher, L.-J. Li, K. Li, and L. Fei-Fei, "ImageNet: A large-scale hierarchical image database," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, 2009, pp. 248–255, doi: 10.1109/CVPR.2009.5206848.
- [8] S. J. Pan and Q. Yang, "A survey on transfer learning," *IEEE Trans. Knowl. Data Eng.*, vol. 22, no. 10, pp. 1345–1359, 2010, doi: 10.1109/TKDE.2009.191.
- [9] K. M. Hosny, M. A. Kassem, and M. M. Foad, "Classification of skin lesions using transfer learning and augmentation with Alex-net," *PLOS ONE*, vol. 14, no. 5, Art. no. e0217293, 2019, doi: 10.1371/journal.pone.0217293.
- [10] K. Simonyan and A. Zisserman, "Very deep convolutional networks for large-scale image recognition," in *Proc. Int. Conf. Learn. Represent. (ICLR)*, 2015, arXiv:1409.1556.
- [11] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, 2016, pp. 770–778, doi: 10.1109/CVPR.2016.90.
- [12] C. Szegedy, V. Vanhoucke, S. Ioffe, J. Shlens, and Z. Wojna, "Rethinking the Inception architecture for computer vision," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, 2016, pp. 2818–2826, doi: 10.1109/CVPR.2016.308.
- [13] M. Tan and Q. V. Le, "EfficientNet: Rethinking model scaling for convolutional neural networks," in *Proc. 36th Int. Conf. Mach. Learn. (ICML)*, PMLR vol. 97, 2019, pp. 6105–6114.
- [14] M. Dildar et al., "Skin cancer detection: A review using deep learning techniques," *Int. J. Environ. Res. Public Health*, vol. 18, no. 10, Art. no. 5479, 2021, doi: 10.3390/ijerph18105479.
- [15] M. Naqvi, S. Q. Gilani, T. Syed, O. Marques, and H.-C. Kim, "Skin cancer detection using deep learning—A review," *Diagnostics*, vol. 13, no. 11, Art. no. 1911, 2023, doi: 10.3390/diagnostics13111911.
- [16] T. J. Brinker et al., "Skin cancer classification using convolutional neural networks: Systematic review," *J. Med. Internet Res.*, vol. 20, no. 10, Art. no. e11936, 2018, doi: 10.2196/11936.

- [17] A. Krizhevsky, I. Sutskever, and G. E. Hinton, "ImageNet classification with deep convolutional neural networks," *Commun. ACM*, vol. 60, no. 6, pp. 84–90, 2017, doi: 10.1145/3065386.
- [18] S. S. Han et al., "Classification of the clinical images for benign and malignant cutaneous tumors using a deep learning algorithm," *J. Invest. Dermatol.*, vol. 138, no. 7, pp. 1529–1538, 2018, doi: 10.1016/j.jid.2018.01.028.
- [19] C. Shorten and T. M. Khoshgoftaar, "A survey on image data augmentation for deep learning," *J. Big Data*, vol. 6, Art. no. 60, 2019, doi: 10.1186/s40537-019-0197-0.
- [20] C. Szegedy et al., "Going deeper with convolutions," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, 2015, pp. 1–9, doi: 10.1109/CVPR.2015.7298594.
- [21] P. Tschandl et al., "Human–computer collaboration for skin cancer recognition," *Nat. Med.*, vol. 26, no. 8, pp. 1229–1234, 2020, doi: 10.1038/s41591-020-0942-0.
- [22] P. Tschandl, C. Rosendahl, and H. Kittler, "The HAM10000 dataset, a large collection of multi-source dermatoscopic images of common pigmented skin lesions," *Sci. Data*, vol. 5, Art. no. 180161, 2018, doi: 10.1038/sdata.2018.161.
- [23] N. C. F. Codella et al., "Skin lesion analysis toward melanoma detection: A challenge at the 2017 International Symposium on Biomedical Imaging (ISBI), hosted by the International Skin Imaging Collaboration (ISIC)," in *Proc. IEEE 15th Int. Symp. Biomed. Imaging (ISBI)*, 2018, pp. 168–172, doi: 10.1109/ISBI.2018.8363547.
- [24] R. Daneshjou et al., "Disparities in dermatology AI performance on a diverse, curated clinical image set," *Sci. Adv.*, vol. 8, no. 32, Art. no. eabq6147, 2022, doi: 10.1126/sciadv.abq6147.
- [25] M. Groh et al., "Evaluating deep neural networks trained on clinical images in dermatology with the Fitzpatrick 17k dataset," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. Workshops (CVPRW)*, 2021, pp. 1820–1828.
- [26] R. R. Selvaraju et al., "Grad-CAM: Visual explanations from deep networks via gradient-based localization," in *Proc. IEEE Int. Conf. Comput. Vis. (ICCV)*, 2017, pp. 618–626, doi: 10.1109/ICCV.2017.74.
- [27] M. Sokolova and G. Lapalme, "A systematic analysis of performance measures for classification tasks," *Inf. Process. Manag.*, vol. 45, no. 4, pp. 427–437, 2009, doi: 10.1016/j.ipm.2009.03.002.
- [28] A. Dosovitskiy et al., "An image is worth 16×16 words: Transformers for image recognition at scale," in *Proc. Int. Conf. Learn. Represent. (ICLR)*, 2021, arXiv:2010.11929.
- [29] M. Goyal, T. Knackstedt, S. Yan, and S. Hassanpour, "Artificial intelligence-based image classification methods for diagnosis of skin cancer: Challenges and opportunities," *Comput. Biol. Med.*, vol. 127, Art. no. 104065, 2020, doi: 10.1016/j.combiomed.2020.104065.
- [30] Y. Wu, B. Chen, A. Zeng, D. Pan, R. Wang, and S. Zhao, "Skin cancer classification with deep learning: A systematic review," *Front. Oncol.*, vol. 12, Art. no. 893972, 2022, doi: 10.3389/fonc.2022.893972.
- [31] J. Yosinski, J. Clune, Y. Bengio, and H. Lipson, "How transferable are features in deep neural networks?" in *Adv. Neural Inf. Process. Syst. (NeurIPS)*, vol. 27, 2014, pp. 3320–3328.
- [32] M. Raghu, C. Zhang, J. Kleinberg, and S. Bengio, "Transfusion: Understanding transfer learning for medical imaging," in *Adv. Neural Inf. Process. Syst. (NeurIPS)*, vol. 32, 2019, pp. 3347–3357.
- [33] G. Huang, Z. Liu, L. van der Maaten, and K. Q. Weinberger, "Densely connected convolutional networks," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, 2017, pp. 2261–2269, doi: 10.1109/CVPR.2017.243.
- [34] M. Sandler, A. Howard, M. Zhu, A. Zhmoginov, and L.-C. Chen, "MobileNetV2: Inverted residuals and linear bottlenecks," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, 2018, pp. 4510–4520, doi: 10.1109/CVPR.2018.00474.
- [35] B. Harangi, "Skin lesion classification with ensembles of deep convolutional neural networks," *J. Biomed. Inform.*, vol. 86, pp. 25–32, 2018, doi: 10.1016/j.jbi.2018.08.006.

- [36] A. Mahbod, G. Schaefer, I. Ellinger, R. Ecker, A. Pitiot, and C. Wang, "Fusing fine-tuned deep features for skin lesion classification," *Comput. Med. Imaging Graph.*, vol. 71, pp. 19–29, 2019.
- [37] N. Gessert, M. Nielsen, M. Shaikh, R. Werner, and A. Schlaefer, "Skin lesion classification using ensembles of multi-resolution EfficientNets with meta data," *MethodsX*, vol. 7, Art. no. 100864, 2020, doi: 10.1016/j.mex.2020.100864.
- [38] M. A. Kassem, K. M. Hosny, and M. M. Fouad, "Skin lesions classification into eight classes for ISIC 2019 using deep convolutional neural network and transfer learning," *IEEE Access*, vol. 8, pp. 114822–114832, 2020, doi: 10.1109/ACCESS.2020.3003890.
- [39] M. A. Marchetti et al., "Results of the 2016 International Skin Imaging Collaboration International Symposium on Biomedical Imaging challenge: Comparison of the accuracy of computer algorithms to dermatologists for the diagnosis of melanoma from dermoscopic images," *J. Am. Acad. Dermatol.*, vol. 78, no. 2, pp. 270–277, 2018, doi: 10.1016/j.jaad.2017.08.016.
- [40] P. Tschandl et al., "Comparison of the accuracy of human readers versus machine-learning algorithms for pigmented skin lesion classification: An open, web-based, international, diagnostic study," *Lancet Oncol.*, vol. 20, no. 7, pp. 938–947, 2019, doi: 10.1016/S1470-2045(19)30333-X.
- [41] Z. Liu et al., "Swin Transformer: Hierarchical vision transformer using shifted windows," in *Proc. IEEE/CVF Int. Conf. Comput. Vis. (ICCV)*, 2021, pp. 10012–10022, doi: 10.1109/ICCV48922.2021.00986.
- [42] P. Li, Y. Xu, Y. Wei, and Y. Yang, "Self-correction for human parsing," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 44, no. 6, pp. 3260–3271, 2022, doi: 10.1109/TPAMI.2020.3048039.
- [43] B. Han et al., "Co-teaching: Robust training of deep neural networks with extremely noisy labels," in *Adv. Neural Inf. Process. Syst. (NeurIPS)*, vol. 31, 2018, pp. 8527–8537.
- [44] C. Guo, G. Pleiss, Y. Sun, and K. Q. Weinberger, "On calibration of modern neural networks," in *Proc. 34th Int. Conf. Mach. Learn. (ICML)*, PMLR vol. 70, 2017, pp. 1321–1330.
- [45] M. Combalia et al., "BCN20000: Dermoscopic lesions in the wild," *arXiv:1908.02288*, 2019.
- [46] V. Rotemberg et al., "A patient-centric dataset of images and metadata for identifying melanomas using clinical context," *Sci. Data*, vol. 8, Art. no. 34, 2021, doi: 10.1038/s41597-021-00815-z.
- [47] B. Cassidy, C. Kendrick, A. Brodzicki, J. Jaworek-Korjakowska, and M. H. Yap, "Analysis of the ISIC image datasets: Usage, benchmarks and recommendations," *Med. Image Anal.*, vol. 75, Art. no. 102305, 2022, doi: 10.1016/j.media.2021.102305.
- [48] C. Barata, M. E. Celebi, and J. S. Marques, "Explainable skin lesion diagnosis using taxonomies," *Pattern Recognit.*, vol. 110, Art. no. 107413, 2021, doi: 10.1016/j.patcog.2020.107413.