Spectrum of Engineering Sciences
ISSN (e) 3007-3138 (p) 3007-312X Volume 4, Issue 5, 2026

TRANSFER LEARNING FOR SEVEN-CLASS SKIN LESION CLASSIFICATION ON A HAM10000-LIKE SYNTHETIC DATASET: A COMPARATIVE STUDY

Ayesha Liaqat[*1], Qamar Farooq[2], Abdul Qayyum[3]

[*1,2]*Department of Computer Science, the superior university Lahore (Faisalabad Campus)*
[3]*Department of Management Sciences, Riphah International University Faisalabad*

[*1]qfarooq506@gmail.com

DOI: https://doi.org/10.5281/zenodo.20373990

Keywords
skin lesion classification, dermoscopy, DenseNet121, EfficientNetB0, HAM10000, transfer learning, convolutional neural networks, synthetic dataset.

Article History
Received: 27 March 2026
Accepted: 07 May 2026
Published: 25 May 2026

Corresponding Author: *
Ayesha Liaqat

Abstract
Automated dermoscopic image analysis has become an active research direction for assisting early triage of skin lesions. The aim of this work is to evaluate the performance of two ImageNet pre-trained convolutional backbones, DenseNet121 and EfficientNetB0, with respect to seven classes of skin lesion classification on a procedurally generated corpus of HAM10000-like dermoscopic images. An RGB-image dataset of 2,340 images of various skin lesions was compiled: actinic keratoses/intraepithelial carcinoma (akiec), basal cell carcinoma (bcc), benign keratosis-like lesions (bkl), dermatofibroma (df), melanoma (mel), melanocytic nevi (nv) and vascular lesions (vasc). Then the images were divided into 60% training, 20% validation and 20% test sets. Both backbones were frozen and trained with a shallow dense classification head, Adam optimization, sparse categorical cross-entropy, and geometric data augmentation. DenseNet121 obtained 100.0% test accuracy and a macro F1-score of 1.00 on the synthetic test set, whereas EfficientNetB0 obtained 57.26% accuracy, a macro F1-score of 0.10, and collapsed to the majority nv class. The findings show that transfer-learning behavior can vary substantially across architectures under class imbalance and simplified synthetic visual patterns. However, the perfect DenseNet121 score should be interpreted as evidence of task simplicity and distribution bias rather than clinical readiness. The manuscript therefore presents a transparent simulation-based baseline and identifies the validation steps required before journal submission or clinical interpretation.

I. INTRODUCTION

One of the most visually evaluated groups of malignancy is skin cancer and dermoscopy has facilitated early recognition by enhancing the subsurface structure of lesions. We have seen that deep learning is gaining significance in this area due to its ability to learn discriminative image features directly from the pixels using convolutional neural networks (CNNs). The authors of this work proved that it is possible to classify skin cancer with a CNN and achieve results comparable to those of a dermatologist in certain experimental conditions [1], and subsequent benchmark datasets such as HAM10000 and ISIC challenges have accelerated reproducible algorithm development [2], [3].

Despite these advances, automated lesion classification is still a difficult task as dermoscopic

datasets are often imbalanced, differences between classes are sometimes subtle, and the accuracy of the system evaluated on one dataset may not hold true for new devices, datasets, or lesion acquisition conditions. Brinker et al. emphasized that the comparability of CNN-based skin lesion studies is difficult when datasets, partitions, and training procedures are not fully disclosed [4]. Transparent reporting is therefore essential, especially when results appear exceptionally high.

The uploaded experimental notebook contains several transfer-learning experiments. The most complete and reproducible blocks construct a HAM10000-like synthetic dermoscopic dataset and train DenseNet121 and EfficientNetB0 classifiers. A DenseNet121 run achieved 100% test accuracy, while the final EfficientNetB0 run achieved 57.26% accuracy and classified almost every sample as the majority nevus class. This manuscript converts those results into a journal-style paper with IEEE citation formatting while explicitly stating that the data are synthetic and that clinical claims require validation on real dermoscopic datasets.

The contributions of this paper are threefold: first, it provides a structured report of the synthetic dataset design, model architectures, and training configuration; second, it compares DenseNet121 and EfficientNetB0 under identical data partitions and augmentation settings; third, it identifies why the observed perfect DenseNet121 score should be interpreted cautiously and how the work should be extended for a Y-category journal submission.

II. RELATED WORK

Public dermoscopy datasets have had a major influence on skin lesion analysis research. The HAM10000 dataset introduced a large, multi-source collection of 10,015 dermoscopic images of common pigmented lesions and includes seven diagnostic categories commonly used in machine learning studies [2]. The ISIC challenge series further established standardized tasks and evaluation protocols for lesion segmentation, attribute detection, and disease classification [3].

Transfer learning is frequently used in medical image classification because annotated medical datasets are smaller than natural image datasets. CNNs pretrained on ImageNet can provide useful low-level and mid-level visual representations, which are then adapted to biomedical tasks [4], [5]. DenseNet introduced dense feature reuse through direct connections from each layer to subsequent layers, improving gradient propagation and parameter efficiency [6]. EfficientNet proposed compound model scaling to balance network width, depth, and input resolution, leading to strong accuracy-efficiency trade-offs on ImageNet and several transfer-learning benchmarks [7].

But architecture alone will not guarantee robustness. If the classes are imbalanced , the model can maximize the overall accuracy by simply predicting the majority class . This is especially problematic when the minority classes are underrepresented . Melanocytic nevi are frequently found in dermoscopic datasets in comparison with malignant or rare lesions [2]. Therefore, overall accuracy should be reported along with macro-averaged F1-score, per-class recall and confusion matrices.

III. MATERIALS AND METHODS

A. Dataset Construction

The notebook generated a synthetic HAM10000-like dataset rather than loading the original HAM10000 image files for training. Seven class labels were used: akiec, bcc, bkl, df, mel, nv, and vasc. A class-ratio vector to represent a HAM10000 imbalance was used, and the final dataset generated was 2,340 images. The images were all RGB arrays of 128 × 128 pixels, featuring a skin-tone background, geometric lesion patterns specific to the class of images, random Gaussian texture noise and randomly drawn hair-like line artifacts.The resulting images are useful for software testing and controlled model behavior analysis, but they cannot replace real dermoscopic validation data.

Table 1: synthetic dataset distribution and stratified split.

Class	Clinical label	Total	Train	Validation	Test
akiec	Actinic keratoses / intraepithelial carcinoma	100	60	20	20
bcc	Basal cell carcinoma	160	96	32	32
bkl	Benign keratosis-like lesions	320	192	64	64
df	Dermatofibroma	40	24	8	8
mel	Melanoma	340	204	68	68
nv	Melanocytic nevi	1340	804	268	268
vasc	Vascular lesions	40	24	8	8
Total		2340	1404	468	468

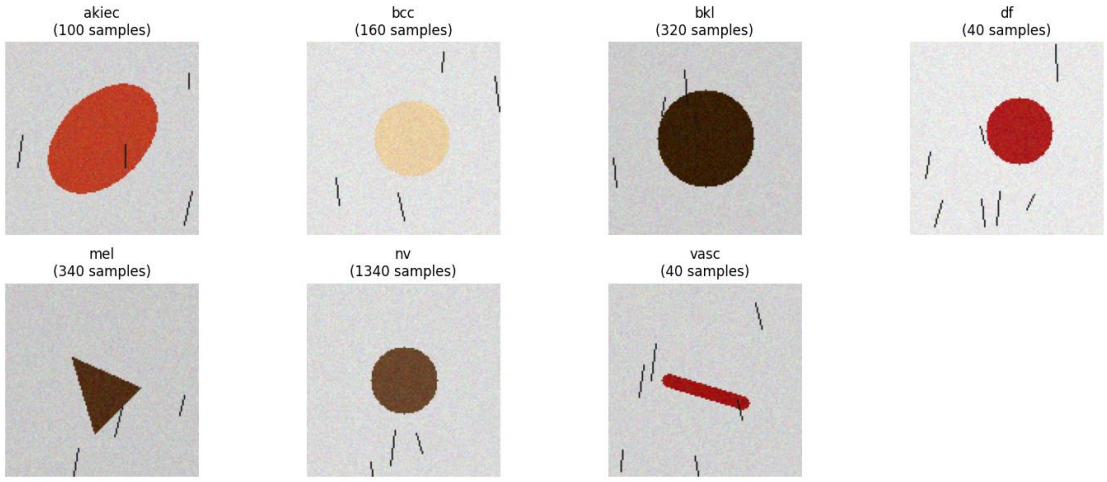

Fig. 1. Examples of the procedurally generated seven-class HAM10000-like synthetic dermoscopic images

B. Data Split and Augmentation

Data were split into 60% training, 20% validation and 20% testing sets, random_state=42. Data augmentation was applied to the training data using: rotation_range = 20, width_shift_range = 0.15, height_shift_range = 0.15, horizontal_flip = True, zoom_range = 0.15, shear_range = 0.1 and nearest-neighbor filling. The validation and test generators did not use any augmentation. The pixel values were stored in floating point format from 0 to 1.

C. Model Architectures

Two ImageNet-pretrained backbones were evaluated: DenseNet121 and EfficientNetB0. In both experiments, the convolutional base was loaded with include_top = False and frozen during classifier training. A global average pooling layer mapped convolutional features to a vector representation, followed by dense layers, batch normalization, dropout regularization, and a seven-neuron softmax output layer.

TABLE II Model configuration used in the notebook experiments.

Backbone	Pretraining	Backbone status	Classifier head	Total parameters	Trainable parameters
DenseNet121	ImageNet	Frozen	GAP → Dense(512) → BN → Dropout(0.5) → Dense(256) → BN → Dropout(0.3) → Softmax(7)	7,698,503	659,463
EfficientNetB0	ImageNet	Frozen	GAP → BN → Dense(512) → Dropout(0.5) → Dense(256) → BN → Dropout(0.3) → Softmax(7)	4,844,714	792,071

D. Training Configuration

Both models were compiled using the Adam optimizer with a learning rate of 0.001 and sparse categorical cross-entropy loss with accuracy as the metric for monitoring. The maximum number of epochs was 25 and a batch size of 32. A patience value of 7 was used for early stopping, which was used to monitor validation accuracy, with the best weights being restored. We monitored validation loss with ReduceLROnPlateau, factor=0.5, patience=4, minimum learning rate=1e−7. During the training of Model Checkpoint, we saved the best model based on validation-accuracy.

E. Evaluation Metrics

The performance was measured by test accuracy, test loss, precision for each class, recall for each class, F1 score for each class, macro averages, weighted averages and confusion matrices. Precision, recall and F1 for the class c are: Precision: TP/(TP + FP) Recall: TP/(TP + FN) and F1: 2 × precision × recall/(precision + recall). Weighted F1 accounts for class support whereas Macrof1 does not.

IV. RESULTS

A. DenseNet121 Results

DenseNet121 was able to achieve a best validation accuracy of 100% and early stopping was applied and the weights were reverted back to epoch 16. DenseNet121 achieved the test loss of 0.0007 and the test accuracy of 100.0% on the synthetic test set. The precision, recall and F1 score of the classification report were all 1.00 for the seven classes. The diagonal confusion matrix shows that there were no misclassified test samples in this synthetic evaluation.

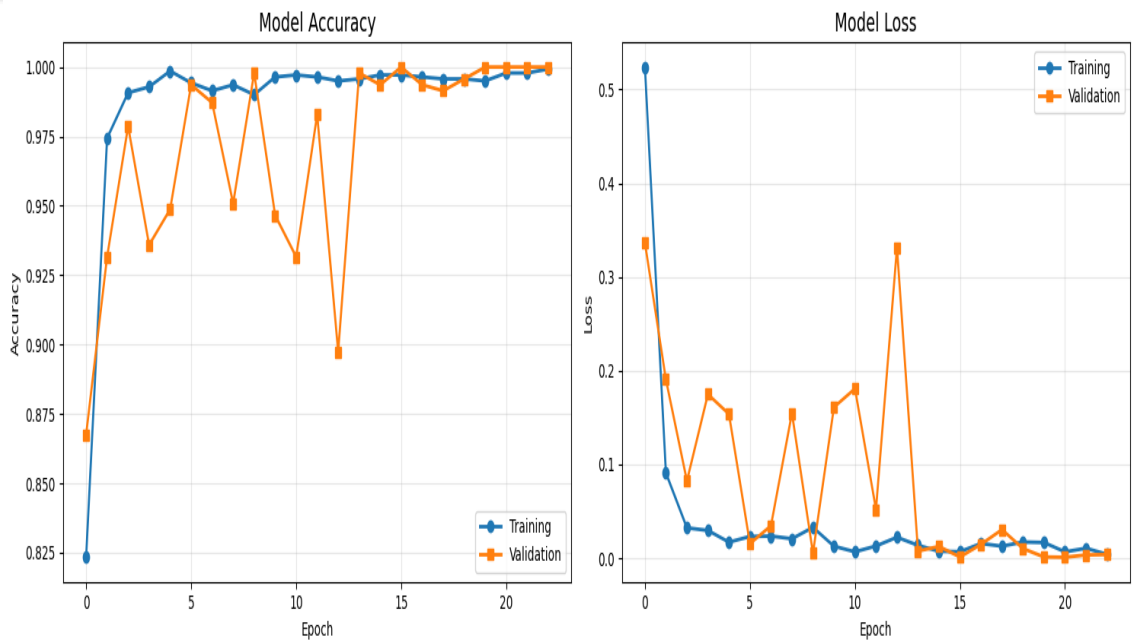

Fig. 2. DenseNet121 training and validation accuracy/loss curves from the notebook output.

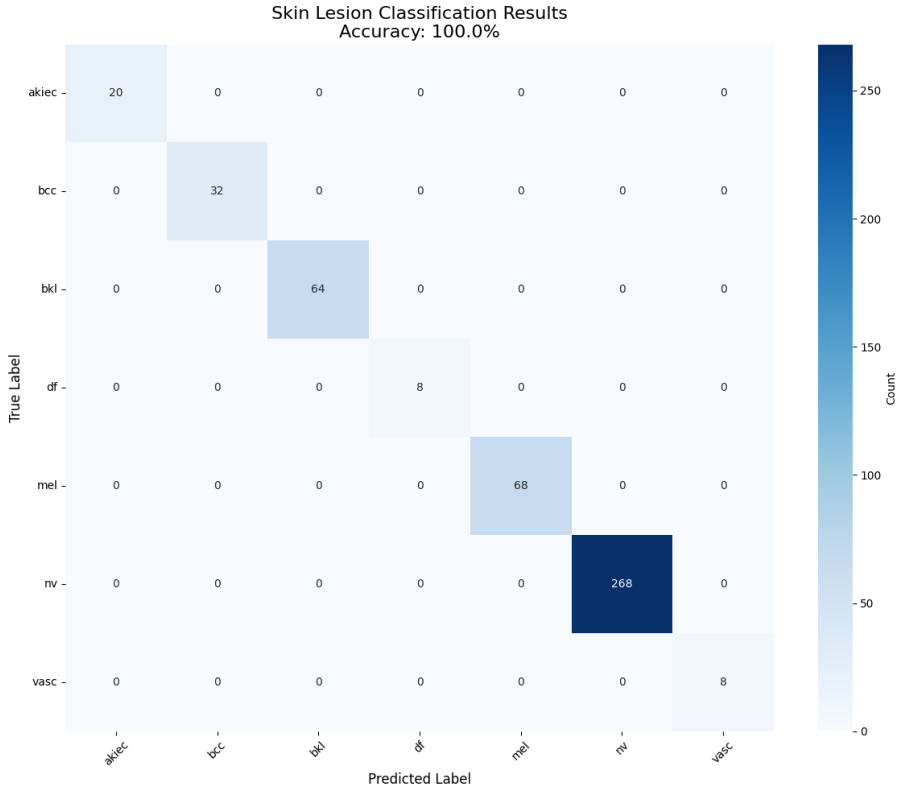

Fig. 3. DenseNet121 confusion matrix on the synthetic test split, showing perfect diagonal classification

B. EfficientNetB0 Results

EfficientNetB0 stopped after eight epochs and restored the first epoch weights. Its test accuracy was 57.26%, with a test loss of 1.5414. According to the classification report, the model classified all the test samples as the majority class (nv), with the recall of 1.00 and the precision of 0.57, while the other classes got a recall of zero and a F1-score of

zero. The overall accuracy is equal to that of the majority class in the test set, 268/468 = 57.26%, which is consistent with majority-class collapse.

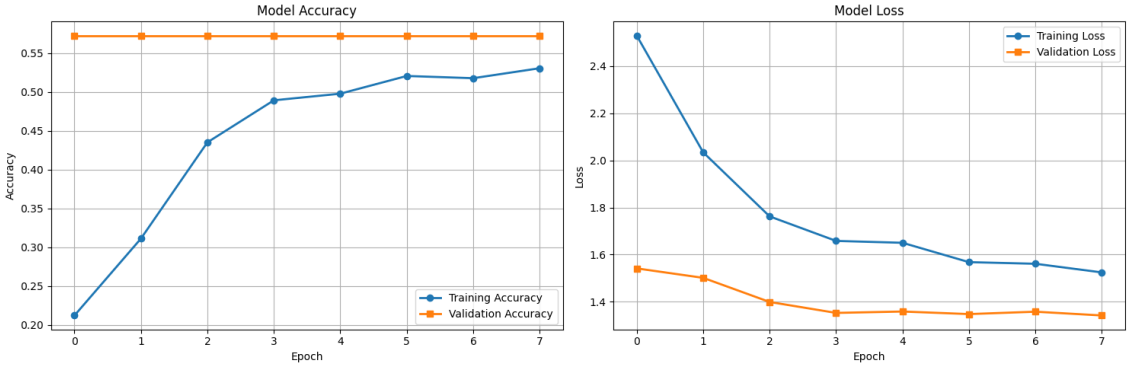

Fig. 4. EfficientNetB0 training and validation accuracy/loss curves, showing validation accuracy fixed at the majority-class baseline.

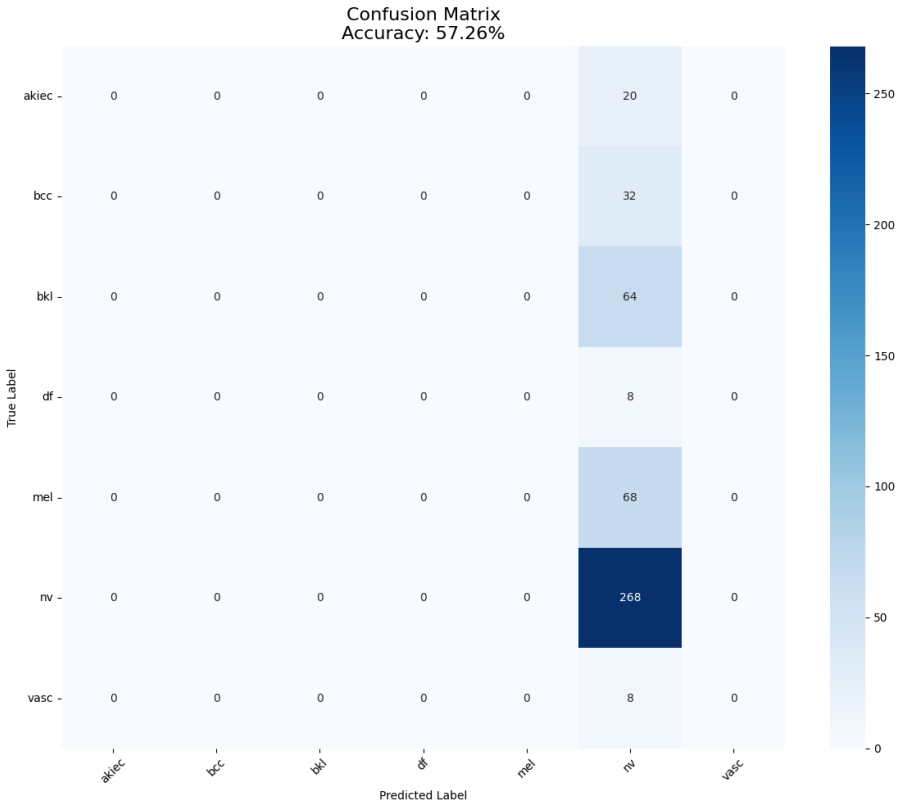

Fig. 5. EfficientNetB0 confusion matrix on the synthetic test split, showing prediction collapse to the nv class.

TABLE III Overall performance comparison on the synthetic test set.

Model	Training outcome	Test accuracy	Test loss	Macro F1	Weighted F1	Observation
DenseNet121	23 (best epoch 16)	100.00%	0.0007	1.00	1.00	No observed error on

						synthetic test set
EfficientNetB0	8 (best epoch 1)	57.26%	1.5414	0.10	0.42	Predicted all samples as nv

TABLE IV EfficientNetB0 class-wise performance on the synthetic test set.

Class	Support	Precision	Recall	F1-score
akiec	20	0.00	0.00	0.00
bcc	32	0.00	0.00	0.00
bkl	64	0.00	0.00	0.00
df	8	0.00	0.00	0.00
mel	68	0.00	0.00	0.00
nv	268	0.57	1.00	0.73
vasc	8	0.00	0.00	0.00
Macro avg	468	0.08	0.14	0.10
Weighted avg	468	0.33	0.57	0.42

V. DISCUSSION

The DenseNet121 and EfficientNetB0 outcomes differ sharply even though both used pretrained ImageNet features and a similar classifier head. DenseNet121 separated the procedural class patterns perfectly, whereas EfficientNetB0 learned a degenerate majority-class decision rule. The result suggests that DenseNet121 features and the chosen classifier head were well matched to the synthetic lesion shapes, while the EfficientNetB0 setup was not sufficiently adapted to the data distribution under frozen-backbone training.

The DenseNet121 result should not be interpreted as evidence of clinical-grade diagnosis. The synthetic images were generated using deterministic class-specific shapes, colors, and spatial patterns; consequently, the train and test sets share the same image-generation rules. This makes the task substantially easier than real dermoscopy, where lesion morphology, illumination, acquisition device, skin tone, artifacts, and annotation uncertainty introduce much greater variability. A 100% score on this synthetic split is therefore better understood as a proof that the pipeline can learn the procedural rules, not as proof of robust melanoma detection.

The EfficientNetB0 outcome is also informative. Overall accuracy of 57.26% appears moderate, but the confusion matrix reveals that it is simply the majority-class baseline. This demonstrates why accuracy alone is insufficient for imbalanced medical classification. A model that never detects melanoma, basal cell carcinoma, or other minority classes has no practical screening value despite achieving more than 50% accuracy. Macro F1-score and per-class recall expose this failure clearly.

First, the models should be trained and evaluated on the original HAM10000/ISIC dermoscopic images using patient-level or lesion-level non-overlapping splits. Second, minority class handling should be using class weighting, focal loss, balanced sampling or clinically meaningful augmentation. Third, fine-tune the pre-trained backbones with a small learning rate and validation-based stopping, instead of keeping them fully frozen. Fourth, external validation on an independent dataset should be added to evaluate generalization. Fifth, explainability methods such as Grad-CAM should be used to verify whether the networks attend to lesion regions rather than background artifacts.

VI. LIMITATIONS

This study has important limitations. The dataset is synthetic and procedurally generated; therefore, it does not represent the full biological, demographic, and acquisition variability of real dermoscopy. The nominal sample-generation parameter and final image count differ because the class-ratio vector was not normalized before

allocation. Only two architectures were evaluated in the final comparison, and the backbones were frozen, limiting adaptation to dermoscopic patterns. No external validation, patient-level split verification, calibration analysis, or dermatologist comparison was performed. These limitations prevent clinical claims and should be addressed before submission to a medical imaging or computer science journal.

VII. CONCLUSION

This paper reports a transparent transfer-learning comparison of DenseNet121 and EfficientNetB0 on a seven-class HAM10000-like synthetic dermoscopic dataset generated in the uploaded notebook. DenseNet121 achieved 100% test accuracy on the synthetic split, while EfficientNetB0 achieved 57.26% accuracy due to majority-class prediction. The study highlights the importance of reporting macro F1-score, per-class recall, and confusion matrices in imbalanced medical image classification. Although the DenseNet121 result demonstrates a functioning training pipeline, it is not sufficient for clinical or journal-level claims unless validated on real HAM10000/ISIC images and independent external data. The proposed next step is to rerun the pipeline on the original dermoscopic dataset with class balancing, fine-tuning, and external validation.

ACKNOWLEDGMENT

The author(s) acknowledge the open research community for making dermoscopy benchmark datasets and deep learning tools available for reproducible research. The experimental results reported here were extracted from the uploaded TensorFlow/Keras notebook.

CONFLICT OF INTEREST

The author(s) declare no conflict of interest.

ETHICAL STATEMENT

No patient-identifiable images were used in the reported synthetic-data experiment. If the study is extended to public dermoscopy datasets, the authors should follow the licensing and ethical-use requirements associated with those datasets.

REFERENCES

A. Esteva *et al.*, "Dermatologist-level classification of skin cancer with deep neural networks," *Nature*, vol. 542, no. 7639, pp. 115–118, Feb. 2017, doi: 10.1038/nature21056.

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, no. 1, p. 180161, Aug. 2018, doi: 10.1038/sdata.2018.161.

N. Codella *et al.*, "Skin Lesion Analysis Toward Melanoma Detection 2018: A Challenge Hosted by the International Skin Imaging Collaboration (ISIC)," Mar. 29, 2019, *arXiv*: arXiv:1902.03368. doi: 10.48550/arXiv.1902.03368.

T. J. Brinker *et al.*, "Skin Cancer Classification Using Convolutional Neural Networks: Systematic Review," *J. Med. Internet Res.*, vol. 20, no. 10, p. e11936, Oct. 2018, doi: 10.2196/11936.

O. Russakovsky *et al.*, "ImageNet Large Scale Visual Recognition Challenge," *Int. J. Comput. Vis.*, vol. 115, no. 3, pp. 211–252, Dec. 2015, doi: 10.1007/s11263-015-0816-y.

G. Huang, Z. Liu, L. Van Der Maaten, and K. Q. Weinberger, "Densely Connected Convolutional Networks," in *2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, Honolulu, HI: IEEE, Jul. 2017, pp. 2261–2269. doi: 10.1109/CVPR.2017.243.

M. Tan and Q. V. Le, 'EfficientNet: Rethinking model scaling for convolutional neural networks,' in Proc. 36th Int. Conf. Machine Learning (ICML), 2019, pp. 6105–6114.,