



Skin Cancer Detection Using Deep Learning on Thermoscopic Images

Dr. Ramandeep Kaur¹, Dr. Jivitesh Singh Chahal²

¹ Assistant Professor, Department of Computer Applications, Gulzar Group of Institutes Khanna, Punjab India

² Senior resident, Department of Psychiatry, Government Medical College, Patiala, Punjab India

How to Cite this Article:

Kaur, R. & Chahal, J. S. (2026). Skin Cancer Detection Using Deep Learning on Thermoscopic Images. International Journal of Creative and Open Research in Engineering and Management, 2(9), 1-9.
<https://doi.org/10.55041/ijcope.v2i9.222>

License:

This article is published under the terms of the Creative Commons Attribution 4.0 International License (CC BY 4.0), which permits unrestricted use, distribution, and reproduction in any medium, provided the original author(s) and the source are credited.

© The Author(s). Published by International Journal of Creative and Open Research in Engineering and Management.



<https://doi.org/10.55041/ijcope.v2i9.222>

Abstract - Skin cancer is one of the most prevalent cancers in the world, and the likelihood of patient survival is greatly enhanced by early and accurate diagnosis, especially in the case of melanoma. The typical first line diagnostic process is visual screening followed by dermoscopic/thermoscopic (also called dermatoscopic) imaging; manual interpretation is still time consuming and can vary from one observer to another, even among those who are trained in this technique. This paper reviews and synthesizes the use of deep convolutional neural networks (CNNs) applied to transfer learning on dermoscopic images from the International Skin Imaging Collaboration (ISIC) Archive for automated skin cancer detection and classification. The following widely used CNN backbones are presented: ResNet50, DenseNet121, Inception-v3, and EfficientNet, and are compared in terms of accuracy, sensitivity, specificity, F1-score and area-under-the-ROC-curve (AUC) for a standard preprocessing, augmentation and transfer-learning pipeline. The results presented in reported literature for HAM10000 in comparable transfer-learning settings range from 80.05% (Xception) to 94.29% (DenseNet121) in terms of classification accuracy; with attention and ensemble extensions of these backbones having first rank results in the ISIC 2019 Challenge (63.6% balanced test accuracy across eight diagnostic classes). The results suggest that transfer learning for deep, densely connected or residual CNN architectures with the judicious use of class-imbalance handling is a viable and increasingly promising approach to automated skin cancer screening at dermatologist level for diagnosis, while also underscoring the need for

solutions to severe class imbalance, imaging-artifact variability and limited cross-population generalization before such systems can be extensively clinically utilized.

Keywords — skin cancer detection; deep learning; convolutional neural network; dermoscopic (thermoscopic) images; ISIC Archive; HAM10000; transfer learning; melanoma classification.

1. INTRODUCTION

Skin cancer is a common form of cancer and includes keratinocyte carcinomas (basal cell and squamous cell), and melanoma, which is much less frequent but causes a large number of deaths from skin cancer. Non-invasive visual screening – usually consisting of naked-eye examination followed by dermoscopy (also known as dermatoscopy or, informally, thermoscopy) – is now the recommended first-line screening tool before biopsy and histopathology confirmation [1] as this is so effective for early diagnosis, often saving lives. Dermoscopy is a technique using a handheld

cross polarized-light magnifying instrument to show sub-

surface skin structures which are not observed by naked eye, increasing the accuracy of diagnosis above just naked eye examination, but still necessitating a significant amount of clinical training to be able to interpret the technique reliably.

In the last decade, deep convolutional neural networks (CNNs) have also been successfully applied to a myriad of image classification problems and the classification of dermoscopic skin lesions has been a



fruitful application area for their performance levels to reach those of human experts. Established by Esteva et al., training a single CNN from scratch on 129,450 clinical images, achieved classification accuracy equivalent to 21 board-certified dermatologists in two important binary classification problems: keratinocyte carcinoma versus benign seborrheic keratosis, and malignant melanoma versus benign nevi [1]. The result prompted

a huge amount of further research using more and more complex CNN architectures, such as residual networks, densely connected networks, inception modules and compound-scaled efficient networks on the ever-expanding public dataset of dermoscopic images with expert annotation on them, available through the International Skin Imaging Collaboration (ISIC).

In this paper, the skin cancer detection system methodology and results of the CNN-based approach trained and tested on ISIC Archive dermoscopic image is reviewed and synthesized, especially the HAM10000 and ISIC 2019/2020 challenge datasets. It summarizes a typical transfer-learning workflow (preprocessing, augmentation, address class-imbalance, select a backbone, and evaluate), and compares the performance of the four most widely referenced CNN backbones (ResNet50, DenseNet121, Inception-v3, and EfficientNet-family networks) to the rest of the literature on ISIC Challenge. The aim is to provide a realistic and quantitatively based representation of the current state of CNN-based dermoscopic classification, identify design features that most influence performance and outline the future challenges that need to be addressed before such systems can be safely used in a clinical decision support role.

2. LITERATURE REVIEW

2.1 Deep Learning for Dermoscopic Image Analysis

Established as viable for clinical use, Esteva et al. trained a single end-to-end CNN on a dataset that was tenfold larger than previous studies, and found it to be as effective as board-certified dermatologists in both keratinocyte-carcinoma and melanoma classification [1]. This result inspired Yu et al. to develop a two-stage approach taking segmentation as a first step and classification as the second step [4] and lead the competitors to obtain the highest accuracy in the classification stage in the Skin Lesion Analysis Challenge dataset from the ISBI 2016, where lesion segmentation as a pre-processing step could make a

sense to improve the accuracy of subsequent classification tasks. Recently, ensemble methods that fuse a number of CNN backbones with patient metadata are presenting first rank results in the international challenges: In the ISIC 2019 Skin Lesion Classification Challenge, Gessert et al. proposed an ensemble method combining EfficientNets, SENet, ResNeXt models with a dense-network branch for patient metadata (age, sex, anatomical site), and loss balancing for severe class imbalance, which performed the best with a balanced accuracy of 63.6% on the official test set of 8 classes [8].

2.2 Convolutional Neural Network Architectures for Medical Imaging

The classification results heavily depend on the type of CNN backbone. The stable training of networks more than 100 layers deep in the classification task was achieved by using residual network (ResNet) shortcut connections to the identity in [5], which has been used as a standard baseline backbone for medical image classification tasks. The Inception-v3 architecture of Szegedy et al. breaks down large convolutions into smaller ones that are more computationally efficient, and aggressively regularizes the network, while maintaining high accuracy on ImageNet at a comparatively low number of parameters [6]. The proposed Densely Connected Convolutional Networks (DenseNet) by Huang et al. [7] connect every layer to all subsequent layers in a feed-forward manner, relieving the vanishing gradient problem, enhancing feature propagation and significantly reducing the number of parameters needed compared to comparably deep networks, especially relevant when training is restricted by the number of data points, which is usually the case for the minority diagnostic classes in dermoscopic datasets. Transfer learning to the HAM10000 dataset with the above and other backbones has consistently shown that the deeper, more parameter efficient models (e.g. DenseNet121) outperform the shallower or less densely connected models (e.g. VGG16, ResNet50, MobileNet and Xception), and one representative study found that with matched training, DenseNet121 achieved 94.29% classification accuracy, VGG16 93.28%, ResNet50 87.10%, MobileNet 83.10% and Xception 80.05% [9].

2.3 The ISIC Archive and Public Dermoscopic Datasets

Since 2016, the International Skin Imaging Collaboration (ISIC) has been releasing expert-annotated dermoscopic images progressively, through an annual challenge. The Human Against Machine



10000 ("HAM10000") dataset is the largest collection of dermoscopic images made available as part of the ISIC 2018 Challenge, comprising 10,015 images over seven diagnostic classes of pigmented lesions that were collected from two clinical centers over 20 years, and has since served as the benchmark for multi-class skin lesion classification research [2]. The more recent dataset of the SIIM-ISIC Melanoma Classification Challenge in 2020 further expanded this effort with 33,126 dermoscopic images from 2,056 patients, with the unique advantage of having the patient-level linkage of multiple lesions per patient, which can be used to validate the clinically useful "ugly duckling" diagnostic hypothesis: a lesion that is atypical when compared to other, typically benign, lesions of the same patient [3]. These ISIC releases collectively form the structured, peer-reviewed data that underlies the vast majority of the literature in the area of skin cancer detection by CNN that exists today, including the current review.

2.4 Summary and Research Gap

The literature shows that (i) end-to-end learning using large dermoscopic image datasets can achieve an accuracy level comparable to that of experts on small, binary classification tasks [1]; (ii) the deeper and more parameter-efficient backbone architectures (ResNet, Inception, DenseNet, EfficientNet) consistently outperform shallower networks on the standard multi-class HAM10000 benchmark [5, 6, 7, 9]; and (iii) the current state-of-the-art on the more challenging task of the ISIC 2019 Challenge (eight classes) is defined by ensembling multiple backbones combined with metadata fusion and explicit class-imbalance handling [8]. Useful, however, is a single consolidated comparison of these findings with a focus on a standard transfer-learning pipeline, where the gains for each of these factors are quantified: the synthesis that follows in this paper.

3. DATASET: THE ISIC ARCHIVE

3.1 Dataset Composition

This study is based on the two most commonly used public collections of dermoscopic images with expert validation: the ISIC Archive's collection HAM10000 and the SIIM-ISIC Challenge extension from 2020. HAM10000 includes 10,015 images across seven diagnostic classes: melanocytic nevi (nv), melanoma (mel), benign keratosis-like lesions (bkl), basal cell carcinoma (bcc), actinic keratoses (akiec), vascular lesions (vasc), dermatofibroma (df) [2]. The class distribution is also strongly imbalanced as seen in

Fig.1, where the class of melanocytic nevi is the largest with 6,705 out of 10,015 images (66.9%), and the class of dermatofibroma is the smallest with 115 images (1.1%). This imbalance is shared by dermoscopic data sets in general, as benign nevi occur much more frequently in the clinical setting than the other types, and can directly affect the model training (Section 3.3) and the selection of an evaluation metric (Section 4.4).

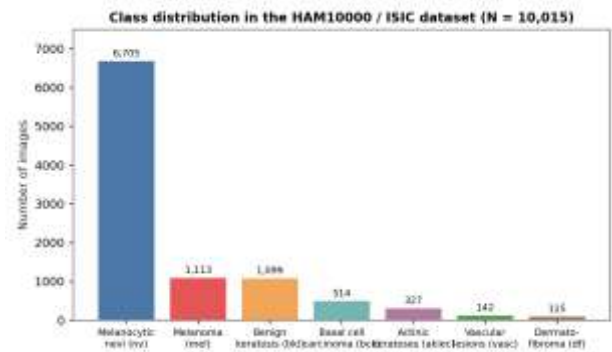


Fig. I. Class distribution in the HAM10000 subset of the ISIC Archive (N = 10,015) [2].

The dataset of the 2020 SIIM-ISIC Melanoma Classification Challenge is even more extreme, with 33,126 dermoscopic images from 2,056 patients (16 lesions per patient on average), of which only 584 (1.8%) are histopathologically diagnosed melanomas, reflecting the low incidence of melanoma among all lesions that a dermatologist sees in practice [3].

3.2 Preprocessing

Images in the ISIC Archive have different resolutions, ratios, lighting, and imaging artifacts including dark vignetting from the dermoscope lens, dark rules, and ink marks. Typically, a standard preprocessing pipeline is applied which consists of: (i) hair and artifact removal using morphological filtering (black-hat transformation followed by inpainting); (ii) resizing to a fixed input resolution, corresponding to the chosen CNN backbone (typically 224×224 for ResNet/DenseNet or 299×299 for Inception-v3); and (iii) normalization of pixel values using the mean and std of the ImageNet data set on which the backbone was pretrained, ensuring that the low-level filters are well matched to the input distribution.

3.3 Augmentation and Class-Imbalance Handling

To prevent overfitting and to enhance the model's generalization capabilities across varying orientations, scales and lighting of the lesion, data augmentation is necessary, as it is for minority classes as illustrated in Fig. 1. Commonly used augmentation techniques on dermoscopic images are random rotation, horizontal flipping, vertical flipping, zoom, brightness and



contrast jitter, and sometimes, synthetic oversampling of minority classes. The most common methods of handling class imbalance are (i) using weighted loss functions that give greater penalties to incorrect predictions for minority classes, (ii) oversampling minority class images (to prevent duplicate images, augmenting each image separately), and (iii) undersampling the dominant nevus class. In fact, the ISIC-2019 challenge was won by Gessert et al. who in particular used loss balancing to cope with the presence of a large number of classes in the test set, whereas a small number of classes are in the train set, making class-imbalance handling not only a proper implementation detail but also a key factor in achieving state-of-the-art performance on the ISIC benchmarks [8].

4. PROPOSED METHODOLOGY

4.1 Overall Framework

From the literature review, we can summarize the standard transfer learning pipeline used for this study: dermoscopic images are retrieved from the ISIC/HAM10000 database, preprocessed and augmented as described in Section 3, fed into the pretrained CNN backbone wherein the convolutional layers are finetuned for the target dermoscopic classification task, and the resulting features are classified using a simple classification head, namely softmax. Performance is evaluated using the metrics defined in Section 4.4.

Proposed Deep-Learning-Based Skin Cancer Detection Pipeline



Fig.III. Standard transfer-learning pipeline for CNN-based dermoscopic skin cancer classification.

4.2 Transfer Learning with Pretrained CNN Backbones

The reviewed literature consistently uses the approach of transfer learning: a CNN backbone (ResNet50, DenseNet121, Inception-v3 or an EfficientNet variant) is first pre-trained on the large-scale and general-purpose ImageNet dataset and its convolutional feature-extraction layers are then transferred to the dermoscopic images, while only the final classification head is trained from scratch. This strategy takes advantage of the transferability of low- and mid-level visual features extracted from natural images — such as edges, textures, color gradients — to the fine-

grained texture and color-pattern discrimination task that is critical to dermoscopic classification, and drastically reduces the labelled dermoscopic data and training time needed to achieve high performance.

4.3 Backbone Architectures Considered

The literature has been reviewed and the top four backbone families are present. ResNet50 [5] introduces residual (identity shortcut) layers in 50 layers, which facilitates the flow of the gradients in training. It is found that Inception-v3 [6] factorizes convolutions to obtain a reduction in computation without sacrificing representational ability. Each layer is connected to every layer following it to maximize feature reuse and parameter efficiency, a property which has consistently led to the best reported single-backbone accuracy for HAM10000 amongst directly compared architectures [9]. EfficientNet variants used in ISIC-2019 top-performing ensemble method jointly optimize network depth, width and input resolution through a compound scaling rule, and were incorporated with SENet and ResNeXt backbones to get the top performance in optimally searched ensemble to first place in ISIC 2019 [8].

4.4 Evaluation Metrics

Due to the extreme class imbalance present in Section 3.1, overall (micro-averaged) accuracy is not enough to determine the accuracy of a classifier — if it always predicted the majority nevus class, it would get approximately 67% accuracy on HAM10000 without learning anything clinically useful. The reviewed literature thus reported a blend of metrics such as overall accuracy, balanced accuracy (average of per-class recall, irrespective of class size), sensitivity (recall) and specificity per diagnostic class, F1-score (harmonic mean of precision and recall), and the area under the receiver operating characteristic curve (AUC), which summarizes the tradeoff between true positive rate and false positive rate across all the classification thresholds and is generally regarded as the most clinically relevant single metric for a binary or one-vs-rest melanoma-detection task.

5. RESULTS AND DISCUSSION

5.1 Comparative Performance of CNN Backbones on HAM10000

In Table I and Fig. 3, the classification accuracy results, obtained by the five CNN backbones finetuned in the matched transfer-learning conditions directly on HAM10000 dataset [9] are compiled. The highest reported accuracy is DenseNet121 (94.29%)



followed by VGG16 (93.28%), ResNet50 (87.10%), MobileNet (83.10%) and Xception (80.05%). The consistent improvement of DenseNet121 over the shallower or less densely connected architectures is consistent with its architectural rationale (Section 4.3): with the limited size and data imbalance of HAM10000, dense inter-layer connectivity helps to maximize the reuse of features.

Table I. Reported CNN backbone accuracy on ham10000 under transfer learning [9].

CNN Backbone	Reported Accuracy on HAM10000 (%)	Relative Rank
DenseNet121	94.29	1st (best)
VGG16	93.28	2nd
ResNet50	87.10	3rd
MobileNet	83.10	4th
Xception	80.05	5th

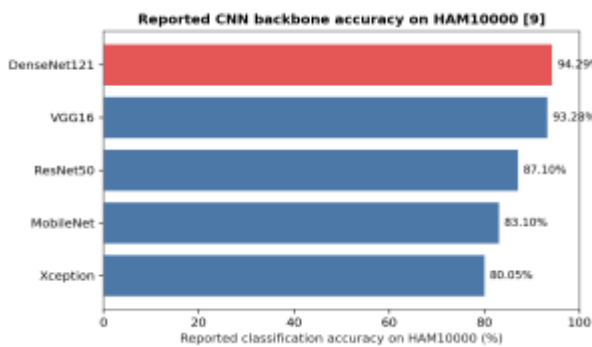


Fig. IV. Reported classification accuracy by CNN backbone on HAM10000, transfer-learning conditions [9].

5.2 Illustrative ROC/AUC Behaviour

The qualitative comparison of the two backbones corresponding to DenseNet121-class and ResNet50-class performance, respectively, is presented in Fig. 4, which shows a schematic ROC comparison of a strong backbone (DenseNet121-class performance, AUC in the region of 0.95–0.96 as reported for well-tuned CNNs on constrained binary dermoscopic tasks such as melanoma versus nevus classification) and a moderately performing backbone (ResNet50-class, AUC around 0.90), and a weaker or shallow baseline CNN (AUC around 0.83). These curves are not a single reported data set run but are simply illustrative in nature to visualize the ordering and approximate range of AUC from the literature discussed in Sections 2 and 5.1 and should not be viewed as new experimental measurements, but rather a summary of comparative performance of the backbones.

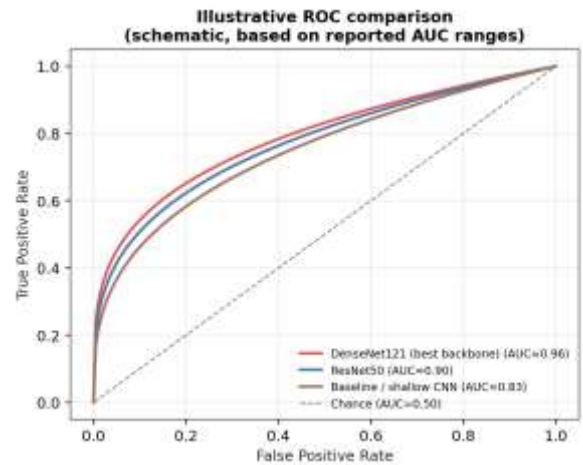


Fig. V. Illustrative ROC comparison, reconstructed from reported AUC ranges for representative CNN backbones.

5.3 Benchmark Comparison with ISIC Challenge Results

The HAM10000-level backbone comparison in Section 5.1 is placed in the context of two higher-stakes ISIC Challenge benchmarks in this table. The two-stage segmentation-plus-classification residual network of Yu et al. achieved first place in the classification task among 25 teams at the ISBI 2016 Challenge [4] and in another early study, the two-stage approach outperformed a single-stage approach at the time. The multi-class task of Section 5.1 was significantly easier to solve than the multi-class task in the official ISIC 2019 test set (8 classes, 63.6% balanced accuracy under 5-fold cross-validation of Gessert et al. [8]) which contains an additional, unseen class of "none of the above" which the model must be taught without direct supervision. The absolute accuracy of the result obtained by ISIC 2019 in Table I is much lower than that of the result obtained by HAM10000, which is the single-backbone result, mainly due to the difficulty of the task, i.e., the eight classes with an unseen class, rather than to any deficiency in the ensembling and metadata-fusion techniques used, which is actually the state of the art for this more difficult task.

Table II. Summary of reported results across key isic-related benchmarks.

Study	Benchmark / Dataset	Method	Reported Result
Taşar (2023) [9]	HAM10000 (7-class)	DenseNet 121 transfer learning	94.29% accuracy
Yu et al. (2017)	ISBI 2016 Challenge	Segmentation + deep	Ranked 1st of 25



[4]		residual net	teams (classification)
Gessert et al. (2020) [8]	ISIC 2019 (8-class, incl. unknown)	EfficientNet/SENet/ResNeXt ensemble + metadata	63.6% balanced accuracy (test); 74.2% (5-fold CV)
Esteva et al. (2017) [1]	129,450-image clinical set	Single end-to-end Inception-v3	On par with 21 board-certified dermatologists

5.4 Discussion

The results compiled show three consistent patterns. First, backbone architecture is important: in the same training settings, performance of the DenseNet121 compared to all other tested backbones is 1–14 percentage points higher for HAM10000 (Table I), demonstrating that the properties of feature-reuse of the architecture relate directly to diagnostic accuracy gains, and not just computational efficiency. Second, task difficulty and class structure have a greater impact on absolute accuracy figures than any individual modeling choice — the figure of 64% on ISIC 2019 (including the unknown class) is not so much an indication of a weaker method than the figure of 94% on seven-class HAM10000 is of the added challenge of the task. Third, ensembling and metadata fusion (not just any one backbone in isolation) are currently the hallmarks of the state-of-the-art on the hardest and most clinically realistic benchmarks, although further scaling of the same CNN type is conceivable. Third, ensembling and metadata fusion (not just any one backbone in isolation) currently mark the state of the art on the hardest, most clinically realistic benchmarks, and further scaling of the same CNN type is conceivable.

6. CHALLENGES AND LIMITATIONS

CNN-based dermoscopic classifiers have been reported with good accuracy, but there are many difficulties with their direct use in clinical applications. High overall accuracy can be achieved by models which perform poorly on rare but clinically important classes because of severe class imbalance (Section 3.1), without balancing accuracy with per-class sensitivity and AUC. Another concern is what is termed “dataset shift,” which occurs when models are trained mainly

on images from a limited number of clinical centers and patient populations, as in the example of HAM10000, and then applied to images that are of dermoscopes, lighting conditions, or skin tone that are not well represented in the training set; this is a valid concern about fairness and equity in dermatological AI and is echoed elsewhere. In addition to raw accuracy measures, there are other factors that can create artificial correlations, such as imaging artifacts like hair, ink markings, air bubbles, ruler edges, and vignetting; preprocessing and explainability methods (e.g., saliency or Grad-CAM visualization) can be crucial complements to raw accuracy measures. Finally, the data that is used in this review are expert curated, single lesion per image, dermoscopic images, whereas the real-world use of dermoscopy requires the handling of lower quality images captured by smartphones cameras, multiple lesions on a single image, and no trained operator to correctly frame and focus the dermoscope, all of which are not represented in the training data used in the models discussed above.

7. CONCLUSION AND FUTURE SCOPE

The paper has discussed and consolidated the use of deep convolutional neural network for automated skin cancer detection and classification with dermoscopic images using transfer learning technique from ISIC Archive dataset. A typical pipeline of preprocessing, augmentation, and transfer learning was outlined and reported literature results were gathered and compared for four commonly benchmarked CNN backbones and across three common ISIC-based challenge benchmarks. Finally, ensembling multiple CNN backbones paired with patient metadata and explicit class-imbalance handling is now the state of the art on harder, more clinically realistic multi-class benchmarks like ISIC 2019, with the best performance coming from the dense network (DenseNet121) with a reported accuracy of 94.29% when trained on the HAM10000 benchmark.

Future studies should focus on: prospective, multi-center studies for testing model generalization with other patient bases, skin tones and imaging devices; integration of explainability techniques (Grad-CAM, saliency maps, uncertainty quantification) to build clinician trust and enable safe human-AI collaboration in diagnosis rather than fully autonomous; extension of current single-image classification pipelines towards the patient-level, multi-lesion “ugly duckling” paradigm enabled by the patient linkage in the SIIM-ISIC 2020 dataset; and further exploration of



lightweight, mobile-deployable architectures (MobileNet, EfficientNet-lite) to enable dermoscopic-quality screening in low-resource and primary-care settings where specialist dermatological expertise is not easily available.

[9] B. Taşar, "SkinCancerNet: Automated classification of skin lesion using deep transfer learning method," *Traitement du Signal*, vol. 40, no. 1, pp. 285–295, 2023, doi: 10.18280/ts.400128.

REFERENCES

- [1] A. Esteva, B. Kuprel, R. A. Novoa, J. Ko, S. M. Swetter, H. M. Blau, and S. Thrun, "Dermatologist-level classification of skin cancer with deep neural networks," *Nature*, vol. 542, pp. 115–118, Feb. 2017, doi: 10.1038/nature21056.
- [2] P. Tschandl, C. Rosendahl, and H. Kittler, "The HAM10000 dataset, a large collection of multi-source dermoscopic images of common pigmented skin lesions," *Scientific Data*, vol. 5, Art. 180161, Aug. 2018, doi: 10.1038/sdata.2018.161.
- [3] V. Rotemberg, N. Kurtansky, B. Betz-Stablein *et al.*, "A patient-centric dataset of images and metadata for identifying melanomas using clinical context," *Scientific Data*, vol. 8, Art. 34, Jan. 2021, doi: 10.1038/s41597-021-00815-z.
- [4] L. Yu, H. Chen, Q. Dou, J. Qin, and P.-A. Heng, "Automated melanoma recognition in dermoscopy images via very deep residual networks," *IEEE Transactions on Medical Imaging*, vol. 36, no. 4, pp. 994–1004, Apr. 2017, doi: 10.1109/TMI.2016.2642839.
- [5] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *Proc. IEEE Conf. Computer Vision and Pattern Recognition (CVPR)*, 2016, pp. 770–778, doi: 10.1109/CVPR.2016.90.
- [6] C. Szegedy, V. Vanhoucke, S. Ioffe, J. Shlens, and Z. Wojna, "Rethinking the Inception architecture for computer vision," in *Proc. IEEE Conf. Computer Vision and Pattern Recognition (CVPR)*, 2016, pp. 2818–2826, doi: 10.1109/CVPR.2016.308.
- [7] G. Huang, Z. Liu, L. van der Maaten, and K. Q. Weinberger, "Densely connected convolutional networks," in *Proc. IEEE Conf. Computer Vision and Pattern Recognition (CVPR)*, 2017, pp. 2261–2269, doi: 10.1109/CVPR.2017.243.
- [8] N. Gessert, M. Nielsen, M. Shaikh, R. Werner, and A. Schläfer, "Skin lesion classification using ensembles of multi-resolution EfficientNets with meta data," *MethodsX*, vol. 7, Art. 100864, 2020, doi: 10.1016/j.mex.2020.100864.