Challenges on Real-World Skin Lesion Classification: Comparing Fine-tuning Strategies for Domain Adaptation using Deep Learning

Tudor Nedelcu, André Carreiro, Francisco Veiga and Maria João M. Vasconcelos
Fraunhofer Portugal AICOS
Porto, Portugal
email: maria.vasconcelos@fraunhofer.pt

***Abstract*—Skin lesion diagnosis is a challenging task even for experimented dermatologists. By using a computer-assisted diagnostic tool, misdiagnosed skin lesions are likely to decrease. Deep neural networks have emerged in recent years due to the increased computational power and their generalization capacity for new data. The major drawback of training a network is that it requires large amounts of data, often difficult to obtain. In this work, we introduce a real-world dataset of single lesions cases composed by clinical images, particularly challenging due to image variations (scale, size, point of view, acquisition device) and data imbalance. To tackle these challenges, we propose a domain adaptation approach by pre-training on large, general, datasets, such as ImageNet and fine-tuning on public datasets of clinical dermatological images. This approach is also compared with one where the target dataset is enriched with samples of public datasets. The overall performance obtained for this real-world dataset was not ideal, with F1-scores below 45%. However, interesting conclusions could be drawn on how fine-tuning generally yields better aggregated results (marginal increase of F1-score), although some specific categories benefit from increased training samples in a merged dataset. These results pave the way for new strategies towards the real-world application of skin lesion classification models, moving forward from more controlled settings, where results are typically impressive, however not yet translatable into practice.**

***Index Terms*—Skin Lesion Classification; Clinical Images; Transfer Learning; Deep Learning**

I. Introduction

There has been a growing interest in Telemedicine and other Information and communications technology (ICT) solutions to improve efficiency [1] and ease the burden on health services, but a significant potential still lies unexplored. Major advances in automatic risk assessment of skin lesions through computer-processed imaging have been recently reported, but most of this work has been conducted solemnly at an academic level and mainly focused on specific parts of the problem. There is a shortage of systems that convert the differently acquired knowledge into reliable decision support tools. Creating an integrated tool with effective practical utility is, thus, critical.

The last decades have seen great improvements concerning computer vision applications for clinical decision support, especially when Machine Learning, and more recently, Deep Learning, came into the picture. In fact, Deep Learning architectures have taken several computer vision tasks to new heights, from which Convolutional Neural Networks (CNN) stand out [2]. The main reason for the popularity of these networks, compared to traditional methods, is that they automatically learn features from images of a specific domain, without any explicit feature engineering. Another reason for their success is the possibility to transfer knowledge acquired for a specific task (resulting in a pre-trained model), to model a different task [2]. Training a CNN [2] from scratch, where the model weights are randomly initialized, requires large amounts of images and repetitive adjustments to the network and its parameters to avoid overfitting the training data [3]. For skin lesion classification, large datasets are a difficult requirement to meet, as the number of images in publicly available datasets is small and expert annotations are expensive [4].

Transfer learning is an alternative to training from scratch that allows to initialize the weights of the layers of a network that we intend to train in a new domain, using weights from a similar network previously trained on data from a different domain. A common practice is to replace the final output layers, where the model decision actually takes place, and freeze all other layers. In this scenario, only the new layers are trained with data from the new domain, keeping the weights of the lower layers fixed (frozen) [5]. This technique is also known as fine-tuning, since only the top layers are trained. In general, the first layers in a CNN learn more generic features, while the last layers learn more specific ones. If both the previously learned and the new domains are similar, fine-tuning the top few layers can be enough. However, if the target domain is considerably different from the source domain (learned by the pre-trained model), we may need to fine-tune the lower layers as well.

In this work, we introduce a new real-world, challenging, dermatological dataset and propose different approaches for training a skin lesion classification model using Deep Neural Networks (DNN) [2]. Firstly, different network architectures are evaluated. Then, different fine-tuning strategies are compared with a baseline where the target dataset is enriched with samples from public datasets containing the same categories. Results are presented and discussed, concerning the impact of the training strategies in classifying individual skin lesion classes, ending with major conclusions and drawing future research lines.

 ISBN: 978-1-61208-916-4

This work integrates a larger project, DermAI, that aims to improve the existing Teledermatology processes between Primary Care Units and Dermatology Services in the Portuguese National Health Service (NHS) for skin lesions referral. Through the usage of Artificial Intelligence and Computer Vision, we envision two major goals: a) to support doctors in Primary Care Units through the development of a mobile app that fosters image acquisition standardization [6] and b) to assist dermatologists in the referral process to book specialist consultations in the Hospital through the adequate prioritization of cases. Improving dermatology consultations' prioritization is particularly relevant in the Portuguese scenario, due to the lack of specialists in the NHS and long waiting lists for this type of consultation. In this research we focus on the second goal of the project towards cases prioritization, firstly on skin lesion classification.

This paper is structured as follows: Section 1 presents the motivation and objectives of this work; Section 2 summarizes background and related work found on the literature; Section 3 provides an overview of the methodology including datasets description, the network architectures studied as well as details on network training; in Section 4, the results and discussion are presented; and finally the conclusions and future work are drawn in Section 5.

II. Background and Related Work

In the last years, several approaches have been studied for using transfer learning in clinical applications. An important aspect of the works found in the literature on skin lesion classification is that the authors use well-established CNN architectures that have achieved excellent performance in large publicly available datasets, such as the ImageNet dataset [7], which consists of natural images of 1000 different categories. Fine-tuning on these pre-trained networks has shown outstanding results in new domains, even with smaller datasets [8] [9], improving both the performance and training times. Lopez et al. [10] proposed a method for skin lesion classification based on dermoscopic images, using a VGG16 network [11] trained for a binary malignant vs. benign classification task. The first four convolutional layers were the result of pre-training the network on ImageNet, whereas the remaining layers were fully trained with new images from the ISIC 2016 dataset [12]. This fine-tuning achieved a sensitivity of 78.66% and precision of 79.74%, which were significantly higher than the top evaluation results for the ISIC 2016 challenge (sensitivity of 50.70% and precision of 63.70%).

Gutman et al. [13] investigated the differences between training a model from scratch, compared with transfer learning and fine-tuning with application to dermatology domain, using EDRA dataset [14]. The model selected was VGG-M [15] with Support Vector Machine (SVM) as a classifier. For transfer learning and fine-tuning, the models were first trained on the Kaggle Retinopathy dataset [16] consisting of retinal images, ImageNet, or both (initially on ImageNet followed by former). The results showed that fine-tuning achieved better results than relying on frozen feature extraction. The models fine-tuned on Retinopathy, or both datasets, led to worse results than when fine-tuning with just ImageNet.

Kawahara et al. [17] trained a linear classifier on features extracted from an AlexNet network [18] pre-trained on ImageNet and fine-tuned on macroscopic images from the Dermofit dataset [19], which classified 10 different skin lesions with high accuracy.

Kawara et al. [20] further used two Inception-V3 [21] pre-trained networks on ImageNet (one for clinical images and one for dermoscopic images), for classification of diagnosis and skin lesion attributes prediction on the EDRA dataset. This work was extended by Nedelcu et al. [22], which pre-trained the networks on the ISIC2019 [23]–[25] dataset for dermoscopic images and Dermofit for clinical images, improving the classification performance.

Mahbod et al. [26] presented an ensemble technique using CNNs for skin lesion classification. The proposed method explores several CNN architectures (AlexNet, VGG16, ResNet-18, and ResNet-101 [27]) pre-trained on ImageNet and fine-tuned with dermoscopic images of skin lesions from the ISIC 2016 [12] and ISIC 2017 [23] datasets. Deep features are extracted from different layers from the different models that are then used to train an SVM. In addition, each pre-trained model is fine-tuned several times with different configurations, boosting the performance of a single architecture and the final results.

In summary, transfer learning has been extensively shown to improve results in different clinical domains, including skin lesion classification. However, this is typically achieved by pre-training in a large, general dataset such as ImageNet, and then fine-tuned on the target dataset, often comprised of dermoscopic images, or, less frequently, macroscopic images but very well standardized in regards to image quality and acquisition conditions. Our proposal to compare different training strategies in a very challenging real-world macroscopic dataset tries to overcome these limitations, including an experiment based on sequential fine-tuning resorting to other datasets, as was done by Gutman et al. [13]. Unfortunately, the results were worsened by using an intermediate dataset, presumably because it was drawn from an entirely different domain (retinal images).

III. Methodology

The main goal of this work is to perform skin lesion classification in a new private dataset, DermAI, consisting of macroscopic and anatomical images from single skin lesions, as shall be explained in Section III-A. Notwithstanding, training a successful classification model from scratch on this dataset is very challenging, especially considering the high number of classes, with a relatively low amount of data for each one (see Table I).

Thus, we explore the potential of fine-tuning on public datasets of related skin lesions, EDRA and Dermofit, further described in Section III-A, and assess how different combinations of these datasets with different characteristics influence the final performance.

 ISBN: 978-1-61208-916-4

As an alternative to transfer learning, we studied the impact of merging the available datasets in the classification performance. Figure 1 illustrates this process, noting that, because the classes considered in each dataset are different, we match DermAI categories with samples in the public datasets, discarding the unmatched categories from EDRA and Dermofit.

Figure 2 illustrates the different fine-tuning strategies under study. The baseline corresponds to training the model for classifying the 13 different skin lesions in DermAI dataset with weights trained from scratch. For the fine-tuning strategies, we fine-tune the DermAI dataset by pre-training on: just ImageNet (a); ImageNet, followed by EDRA (b); and finally, ImageNet, followed by EDRA and then by Dermofit (c). The next subsections describe the datasets used, the network architecture and parameters used for training and validation.

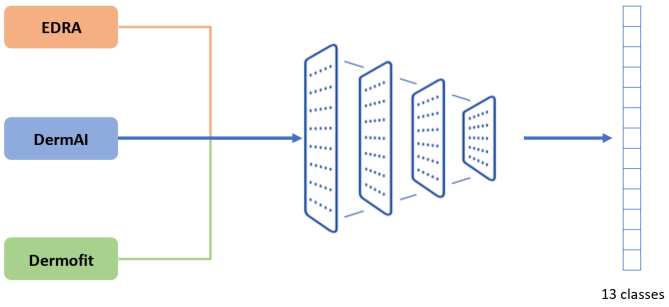


Figure 1. Training strategy based on a merged dataset. The goal is to classify the skin lesions available in DermAI. Thus, there is a previous mapping of the classes from the EDRA and Dermofit datasets, where unmatched class examples are dropped.

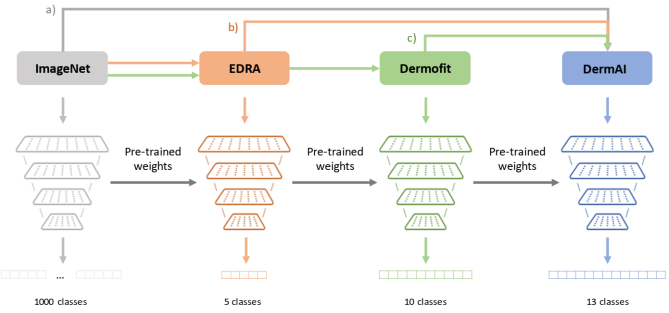


Figure 2. Training strategy based on transfer learning. The baseline considers only the training of DermAI from scratch, and then fine-tuning on DermAI is done sequentially and iteratively by pre-training on a) just ImageNet; b) ImageNet followed by EDRA; c) ImageNet followed by Edra and Dermofit.

A. Datasets

1) DermAI: With a larger goal of building a prioritization model for dermatological referrals, the authors had access to retrospective data from the Portuguese National Health System related to the referral requests from Local Health Care Units for the first Dermatology Hospital consultation. The cases correspond to requests that occurred between the implementation of this referral procedure for dermatological requests in 2013 to the end of February 2020, before the onset of the COVID-19 pandemic in Portugal. Since this data is retrospective it is not possible to publicly release it, due to confidentiality questions and impossibility to get consent from the patients.

After analyzing the available data, and together with a group of dermatologists, it was decided to proceed with a subset of 3430 cases corresponding to single lesions from 13 distinct differential diagnoses. The average age (and standard deviation) of the individuals correspond to 55.75 ± 22.21, and regarding the sex, there are 1422 Male cases, and 2042 Female instances. The distribution of cases in relation to the differential diagnosis provided by dermatologists is presented in Table I. Although teledermatology guidelines [28] recommend the acquisition of macroscopic images, in practice this is not always the case. Each case in the DermAI dataset has an associated clinical image: close-up/macroscopic, or anatomical. The dataset contains 3134 macroscopic images and 296 anatomical ones, as described in Table I. The type of images was selected by examining each type of image by the authors. In particular cases such as in lesions present in hands, arms, feet, or faces, it is difficult to differentiate between the macroscopic and anatomic labels since they look similar (close enough to evaluate the lesion, but wide enough to distinguish the anatomical site). Thus, it was decided to merge these modalities to train a DNN model. By merging the datasets, the number of samples is increased, which is beneficial for a smaller dataset such as this one.

Furthermore, as the authors had no access to biopsy results, in order to have increased confidence in the data, the authors have previously asked a set of dermatologists to review and validate a considerable subset of the test set.

2) EDRA: This is a public dataset [14] [20] that contains both clinical and dermoscopic images, as well as patient metadata. Clinical images are less standardized when compared with dermoscopic ones, meaning they are taken at various fields of view, and can also contain image artefacts such as rulers or other markers. Patient metadata includes other types of information, such as patient gender and lesion location. The dataset contains a total of 1101 cases distributed into 5 different categories: Seborrheic Keratosis, Miscellaneous (dermatofibroma, lentigo, melanosis, vascular lesion, miscellaneous), Nevus (blue, Clark, combined, congenital, dermal, recurrent, reed), Basal Cell Carcinoma and Melanoma. These classifications were assigned by a dermatologist, and the case distribution can be observed in Table II, where, for the scope of this work, only clinical images were considered. Moreover, some examples are presented in Figure 3.

3) Dermofit: Another public dataset is the Dermofit digital image database [19], which consists of 1300 high-quality color skin lesion images taken with standard cameras. The lesions belong to 10 different categories with 819 benign and 481 carcinogenic images, annotated in individual diagnostic classes (Table II). This dataset contains only close-up/macroscopic images and is the most standardized compared with the previous datasets, as it can be observed in Figure 3.

 ISBN: 978-1-61208-916-4

TABLE I
DERMAI DIFFERENTIAL DIAGNOSIS DATASET DISTRIBUTION.

Class	Differential diagnosis	Mac.	Anat.	Total
1 SebKer	Seborrheic Keratosis	1125	61	1186
2 ActKer	Actinic Keratosis	442	77	519
3 Nev	Nevus, Non-neoplastic	561	57	618
4 MolCont	Molluscum Contagiosum	50	21	71
5 Haem	Haemangioma	66	4	70
6 UncNeop	Neoplasm Unc. Behavior	233	13	246
7 Drmfib	Dermatofibroma	135	6	141
8 SLent	Solar Lentigo	45	3	48
9 PenFib	Pendulum Fibroma	99	16	115
10 VWart	Viral Warts	167	25	192
11 OtMalNeop	Other Malignant Neoplasm	108	8	116
12 BCC	Basal Cell Carcinoma	53	3	56
13 MM	Malignant Melanoma	50	2	52
	Total	3134	296	3430

TABLE II
EDRA AND DERMOFIT DIAGNOSIS DATASET DISTRIBUTION.

EDRA

Diagnosis	Total
Seborrheic Keratosis	45
Miscellaneous	97
Nevus	575
Basal Cell Carcinoma	42
Melanoma	252
Total	1101

Dermofit

Diagnosis	Total
Seborrheic Keratosis	257
Actinic Keratosis	45
Melanocytic Nevus	331
Haemangioma	97
Pyogenic Granuloma	24
Dermatofibroma	65
Intraepithelial Carcinoma	78
Squamous Cell Carcinoma	88
Basal Cell Carcinoma	239
Malignant Melanoma	76
Total	1300

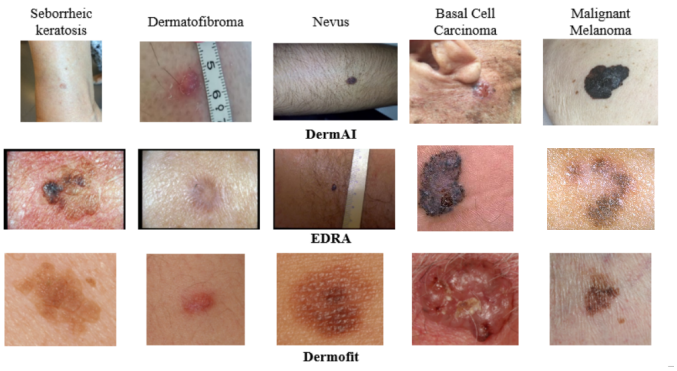


Figure 3. Illustrative examples of lesions from DermAI, EDRA and Dermofit datasets.

B. Merged Dataset

Given the unbalanced nature of the DermAI dataset, with some of the underrepresented classes being available in public datasets such as EDRA and Dermofit, we designed an experiment where such categories would be enriched by samples from those data sources. Examples of such classes, present in both EDRA and Dermofit, that can be joined into DermAI are Seborrheic Keratosis (SebKer), Nevus (Nev), Basal Cell Carcinoma (BCC), and Malignant Melanoma (MM). Additionally, from Dermofit we joined samples from Actinic Keratosis (ActKer) and Haemangioma (Haem). Furthermore, due to its low number of available samples, as well as its typical prioritization assessment in referrals to specialists, we considered examples from Pyogenic Granuloma into DermAI's class of Neoplasm of Uncertain Behavior (UncNeop), and images from Interepithelial Carcinoma and Squamous Cell Carcinoma into DermAI's category of Other Malignant Neoplasms (OtMalNeop). From the EDRA dataset, we extracted samples from the Miscellaneous class corresponding to Dermatofibroma (Drmfib) and Solar Lentigo (SLent) categories. The remaining classes in DermAI remained unchanged.

C. Network Architecture

Three different networks were studied, MobileNet-V2 [29], ResNet50 [30] and EfficientNet-B3 [31]. The MobileNet-V2 is one of the most adopted network for edge devices and is based on an inverted residual structure where the shortcut connections are between the thin bottleneck layers2 [29]. The ResNet50 is a residual network where the residual blocks make it easier to optimize, gaining accuracy from considerably increased depth [30]. The EfficientNet is a group of networks developed based on the network scaling (depths, width and resolution) [31]. An analysis performed on scaling ResNet and MobileNet networks has shown an increase on classification prediction on ImageNet. Although EfficientNet-B3 [31] was shown to surpass the other networks on ImageNet dataset, we also assess their performance on DermAI dataset for skin lesion classification.

On top of each architecture, a few layers are included for the final prediction. A fully connected (dense) layer is applied on top of the extracted feature map (EfficientNet-B3 $10 \times 10 \times 1536$, MobileNet-V2 $10 \times 10 \times 1280$, ResNet $10 \times 10 \times 2080$), generating a number of channels related to the number of classes to predict. For the DermAI dataset, the shape of this layer is $10 \times 10 \times 13$. For dimensionality reduction, the Global Average Pooling method is applied ($1 \times 1 \times 13$) since is known to reduce overfitting [32]. The final output is obtained by using the *softmax* activation function.

The input of the network consists of images of size 300×300. Since the images from the datasets have different shapes, we resize the images to the desired shape using the nearest neighbor method.

D. Network Training

The data is split into Train set and Test set with a ratio of 80:20 considering a stratified distribution of the classes.

The network is trained using the weights pre-trained on ImageNet (or from scratch with random initialization). The frozen block approach is adopted for better results [33]. Each block is trained for 3 epochs using a learning rate of 10^{-4} for the top layer and 10^{-5} for the rest of the blocks. Considering the EfficientNet-B3 architecture, 7 blocks are used for training (the classification block and other 6 modules). Adam is used as an optimizer, and the considered loss is the categorical cross-entropy.

A similar approach is followed for the MobileNet-V2 and ResNet50 networks (pre-trained on ImageNet), where

 ISBN: 978-1-61208-916-4

TABLE III
AVERAGE METRICS SCORE FOR DIFFERENT NETWORKS, AFTER PRE-TRAINING WITH IMAGENET (IN %).

Experiments	Number of Parameters	Average Accuracy	Weighted F1	Macro F1
MobileNet-V2	2.3M	14.43	15.91	9.59
ResNet50	23M	43.00	42.67	27.07
EfficientNet-B3	12M	42.71	44.04	28.65

the difference lies in the block mapping since the network architectures are different. For ResNet50, the classification block and 4 modules are considered, and for MobileNet-V2 the last 11 blocks. The blocks from MobileNet-V2 are grouped because of the residual connections as: block A - block 16; block B - block 13, 14, 15; block C - block 10, 11, 12; block D - block 6, 7, 8, 9.

To mitigate possible overfitting issues due to imbalanced data, we considered stratified batches, where the batch size was chosen to match the number of classes for each data set (5, 10, and 13 for EDRA, Dermofit, and DermAI, respectively). This results in oversampling of the classes with fewer examples. Additionally, we augmented the training data using simple techniques: rotation in the range of $[1, 30]$ degrees, horizontal flip, zooming in the range $[0, 0.2]$, width shift in the range $[0, 0.1]$, and brightness in the range $[0.2, 0.8]$.

IV. RESULTS AND DISCUSSION

The average metrics obtained for the three networks tested (with weights from ImageNet and fine-tuned on DermAI dataset) are presented in Table III. One can observe that MobileNet-V2 is performing poorly, with the network failing to provide acceptable performances. The results obtained using ResNet50 are similar to EfficientNet-B3, although more parameters are used. EfficientNet-B3 uses approximately 12 million parameters, whereas the ResNet50 uses almost double that number (23 million). Therefore, we chose to proceed with the experiments considering the EfficientNet-B3 network.

Table IV summarizes the overall aggregated results for the different considered approaches based on the chosen EfficientNet-B3 network architecture: training from scratch (o), using a merged version of the dataset (x), or different fine-tuning strategies (a-c).

The first experiment was to train the target dataset (DermAI) from scratch, to assess the importance of using pre-trained weights, even in general domains like ImageNet. The very poor results (Accuracy under 15% and F1-score macro under 3%) confirm this, and although we do not show the confusion matrix due to space constraints, it was observed a clear bias towards classifying most samples as Nevus (Nev) and Haemangioma (Haem), the second most and fourth least represented classes.

On the other hand, using models pre-trained on ImageNet drastically improved the results, even though these are still under what is expected from a clinical decision support tool in production. Comparing the averaged metrics, we can see that using a merged version of the training dataset (merging samples from EDRA and Dermofit into DermAI, where a match could be found between classes), returned a slightly lower accuracy and F1-score (weighted and macro) than when fine-tuning is used, be it just on DermAI, or sequentially with EDRA and Dermofit. Nonetheless, using a merged dataset returned some interesting results for particular categories, as can be seen in Table V. For instance, for the Haemangioma class, this approach was the one correctly classifying more samples, which may indicate that in some specific cases, there are discriminative features that can be more easily learned in the same learning process, although getting lost in the iterative process of sequential fine-tuning.

The results for fine-tuning strategies (a), (b), and (c) can be found on Tables VI to VIII. Through the analysis of the previous metrics and confusion matrices (where the predicted labels are on the abscissa and true labels are on the ordinate), comparing the different fine-tuning strategies, interesting findings can be highlighted for the individual categories of skin lesions.

TABLE IV
AVERAGE METRICS SCORE FOR DIFFERENT EXPERIMENTS (IN %).

Experiments	Aver. Acc.	Weight. F1	Macro F1
o) Training from scratch	14.28	5.52	2.53
x) Pre-train. ImageNet, merged Dataset	42.56	43.34	28.60
a) Pre-train. ImageNet	42.71	44.04	28.65
b) Pre-train. ImageNet and EDRA	43.73	44.17	30.09
c) Pre-train. ImageNet, EDRA, Dermofit	43.44	44.41	28.80

Regarding Seborrheic Keratosis (SebKer), which is the most represented class in the DermAI dataset, we report an F1-score of approximately 56% for pre-training on just ImageNet (a) and marginally higher for fine-tuning on EDRA (b) and additionally on Dermofit (c), with (b) and (c) revealing slightly lower sensitivity and higher precision. This class is well represented in the three considered datasets, and as expected, its classification was improved through more fine-tuning, although some misclassifications still happen, especially with the categories of Actinic Keratosis (ActKer), Nevus (Nev), and Neoplasm of Uncertain Behavior (UncNeop).

For the Actinic Keratosis (ActKer), also present in Dermofit but not in EDRA, the model returned F1-scores of approximately 52%, 56%, and 54% for strategies a), b) and c), respectively. This increase is due to the increase in sensitivity in strategies b) and c). This is especially interesting for b), given that this category is absent from the EDRA dataset, which might indicate that other categories from that dataset may present similar features which help to discriminate Actinic Keratosis. Typical misclassifications of this class include Seborrheic Keratosis (SebKer), Other Malignant Neoplasms (OtMalNeop), and Neoplasm of Uncertain Behavior (UncNeop).

Concerning the Nevus class (Nev), the second most represented category in DermAI and present in the 3 datasets (although EDRA comprises Melanocytic Nevus), the obtained F1-score results were very similar for all fine-tuning strategies

 ISBN: 978-1-61208-916-4

TABLE V
RESULTING METRICS WHEN PRE-TRAINING WITH IMAGENET AND FINE-TUNING ON MERGED DATASET, AND CORRESPONDING CONFUSION MATRIX.

Classes	Sens.	Prec.	F1
1 SebKer	53.59	66.84	59.48
2 ActKer	63.46	57.89	60.55
3 Nev	18.55	56.10	27.88
4 MolCont	42.86	27.27	33.33
5 Haem	21.43	21.43	21.43
6 UncNeop	18.37	11.11	13.85
7 Drmfib	53.57	31.25	39.47
8 SLent	10.00	8.33	9.09
9 PenFib	43.48	33.33	37.74
10 VWart	66.67	38.81	49.06
11 OtMalNeop	26.09	16.22	20.00
12 BCC	0.00	0.00	0.00
13 MM	0.00	0.00	0.00

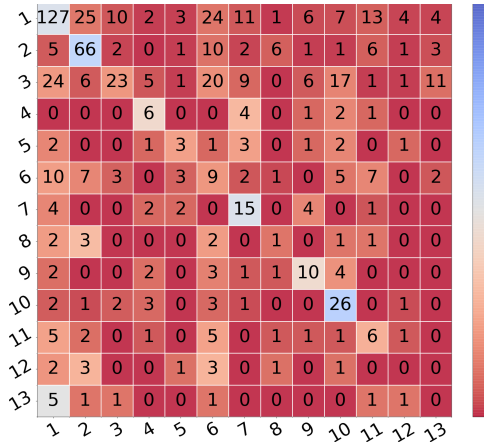
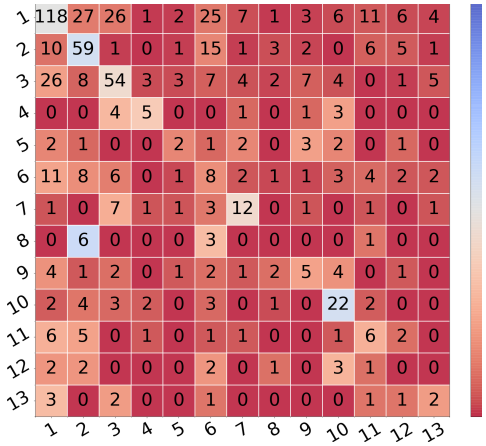


TABLE VI
RESULTING METRICS WHEN PRE-TRAINING WITH IMAGENET AND FINE-TUNING ON DERMAI, AND CORRESPONDING CONFUSION MATRIX.

Classes	Sens.	Prec.	F1
1 SebKer	49.79	63.78	55.92
2 ActKer	56.73	48.76	52.44
3 Nev	43.55	51.43	47.16
4 MolCont	35.71	38.46	37.04
5 Haem	14.29	18.18	16.00
6 UncNeop	16.33	11.27	13.33
7 Drmfib	42.86	38.71	40.68
8 SLent	0.00	0.00	0.00
9 PenFib	21.74	21.74	21.74
10 VWart	56.41	45.83	50.57
11 OtMalNeop	26.09	18.18	21.43
12 BCC	0.00	0.00	0.00
13 MM	20.00	13.33	16.00



(47%, 47%, and 48%, for a), b), c), respectively). However, looking at the sensitivity scores (44%, 46%, 48%), we can see that fine-tuning with additional dermatological datasets where examples of that class were present, considerably improved this important metric. Misclassifications for this category are biased towards Sebhorreic keratosis (SebKer), Neoplasm of Uncertain Behavior (UncNeop), and Dermatofibroma (Drmfib).

Analyzing the *Molluscum Contagiosum* class (absent from EDRA and Dermofit), we can observe F1-scores of approximately 37%, 43% and 38% for strategies a), b) and c). As for ActKer, fine-tuning with data with similar features, even though for different skin lesion categories, helps to better generalize when the number of samples is low, as is the case for this class.

A challenging category is Haemangioma (Haem), given the low amount of data. The F1-scores of 16%, 9% and 15% for strategies a), b), and c), does not support the advantage of using fine-tuning in this particular category, even though it is also present in Dermofit. Analyzing the erroneous classifications, these were more or less evenly distributed among different classes, such as Nevus, Pendulum Fibroma, Dermatofibroma, Sebhorreic Keratosis, and Viral Warts. To better understand these results, we analyzed these images more closely, concluding that, besides a low number of samples, they are very different amongst themselves (e.g., different body regions), making the model task more difficult.

Regarding the Dermatofibroma (Drmfib) category, which also exists in Dermofit and EDRA, the model returned F1-scores of 41%, 36%, and 33% for the three fine-tuning strategies (a), b), c), respectively). Although the differences seem considerable, given the small number of test cases, this translates to a difference of two correctly classified cases between a) and b), and one case between a) and c). Most misclassifications classify the samples as Nevus, and on a lesser extent with Neoplasm of uncertain behavior.

For Solar Lentigo (SLent), which has very few samples (under 50), and only exists in EDRA within the Miscellaneous class, returned poor results, as can be seen by the F1-scores of 0%, 17% and 0% for strategies a), b) and c), respectively. Again, as the difference might seem considerable at first, this corresponds to 0, 2, and 0 correctly classified samples in the test set. This category is expectedly difficult due to its low availability, with the model misclassifying these samples mostly with Actinic and Seborrheic Keratoses.

Concerning Pendulum Fibroma (PendFib), a class only present in DermAI, the results for F1-score for the fine-tuning strategies a), b), c) were 22%, 26%, and 24%, respectively. Even if only marginally, and given the low number of available images, pre-training seems to improve the results, especially

 ISBN: 978-1-61208-916-4

TABLE VII
RESULTING METRICS WHEN PRE-TRAINING WITH IMAGENET AND FINE-TUNING ON EDRA AND DERMAI, AND CORRESPONDING CONFUSION MATRIX.

Classes	Sens.	Prec.	F1
1 SebKer	47.68	69.75	56.64
2 ActKer	67.31	47.30	55.56
3 Nev	45.97	48.31	47.11
4 MolCont	42.86	42.86	42.86
5 Haem	7.14	11.11	8.70
6 UncNeop	4.08	5.13	4.55
7 Drmfib	50.00	28.00	35.90
8 SLent	20.00	15.38	17.39
9 PenFib	26.09	25.00	25.53
10 VWart	53.85	42.86	47.73
11 OtMalNeop	17.39	12.50	14.55
12 BCC	9.09	6.25	7.41
13 MM	30.00	25.00	27.27

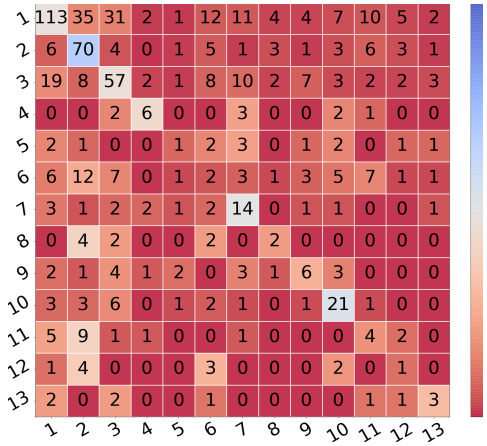
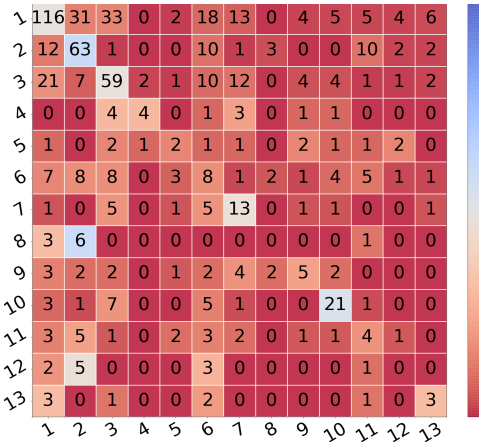


TABLE VIII
RESULTING METRICS WHEN PRE-TRAINING WITH IMAGENET AND FINE-TUNING ON EDRA, DERMOFIT, AND DERMAI, AS WELL AS THE CORRESPONDING CONFUSION MATRIX.

Classes	Sens.	Prec.	F1
1 SebKer	48.95	66.29	56.31
2 ActKer	60.58	49.22	54.31
3 Nev	47.58	47.97	47.77
4 MolCont	28.57	57.14	38.10
5 Haem	14.29	16.67	15.38
6 UncNeop	16.33	11.76	13.68
7 Drmfib	46.43	25.49	32.91
8 SLent	0.00	0.00	0.00
9 PenFib	21.74	26.32	23.81
10 VWart	53.85	52.50	53.16
11 OtMalNeop	17.39	13.33	15.09
12 BCC	0.00	0.00	0.00
13 MM	30.00	20.00	24.00



on sensitivity (22% in a) to 26% in b)). Misclassifications for this class do not reveal a clear trend, spanning several different categories.

The category of Viral Warts (VWart) is also present only in DermAI. The F1-score results of 51%, 48%, and 53% for strategies a), b), c), respectively reveal that this is one of the best-classified categories, and even though Dermofit does not present this category, pre-training with this dataset marginally improves the performance, especially regarding the precision (53%). Most misclassifications fall in the categories of Nevus, and Seborrheic and Actinic Keratoses, the most represented classes, which might support this observed bias, even though class-balancing methods were explored.

The category representing Neoplasm of Uncertain Behavior (UncNeop), as the name suggests, shows a higher variability within its samples. Thus, it is expected that this class is more challenging to classify, which is reflected in lower F1-scores of 13%, 5%, and 14% for fine-tuning strategies a), b), c), respectively. The drop for strategy b) might be a reflection of the under-representation of categories with similar features in the EDRA dataset. We recall that we have included the Pyogenic Granuloma cases into this category due to its low number of available samples and its clinical manifestations resulting in highly variable prioritization in case of referral, which is consistent with the UncNeop category. Since Dermofit also considers some examples of Pyogenic Granuloma, this might explain the marginal improvement of strategy c). Regarding misclassifications, these are mostly distributed among Seborrheic and Actinic Keratoses, Nevus, and Other Malignant Neoplasms, which is understandable as these are some of the classes with more diversity in their clinical manifestations.

Moving from the benign classes to the malignant ones, we start with Other Malignant Neoplasms (OtMalNeop), a broader category that includes the Intraepithelial Carcinoma and Squamous Cell Carcinoma, both present in the Dermofit dataset. Given the high case diversity, not supported by a sufficient number of data samples, lower results are somewhat expected, as observed in the F1-scores of approximately 21%, 15%, and 15% for the strategies a), b), and c), respectively. The drop in performance when pre-training with different datasets suggest that even though some of the learned features may be more general, this does not always imply better robustness for more naturally diverse categories, which can benefit from learning more freely on the available specific samples. Not surprisingly, both from a clinical perspective and due to their representativity in the dataset, most misclassifications fall in the Actinic and Sebhorreic Keratoses classes.

For Basal Cell Carcinoma (BCC), a class that is present in the three datasets, although with a low number of samples

 ISBN: 978-1-61208-916-4

in EDRA and DermAI, the model was only able to correctly classify one sample using fine-tuning strategy b) and none for the remaining strategies. BCC can have different biological and clinical manifestations [34], which by itself makes this class more complex to classify, especially with a lower sample availability. Nonetheless, the results motivated a deeper analysis of the corresponding images, and it was concluded that, for the DermAI dataset, the vast majority of examples in this class were anatomical images (mostly faces), whereas the images on EDRA and Dermofit were macroscopic images, centered on a single lesion. This mismatch between datasets, and even between partitions (train and test sets), together with the anticipated complexity, explains the poor performance for this category.

Finally, regarding the Malignant Melanoma (MM) class, it has examples in the three datasets, although more represented in EDRA. The classification performance resulted in F1-scores of 16%, 27%, and 24% for fine-tuning strategies a), b), and c), respectively. However, in absolute terms, given the low number of test samples (10), this corresponds to 2, 3, and 3 correctly classified instances. In terms of sensitivity, the values were 20%, 30%, and 30%, which hints at the positive impact of fine-tuning for this category where sensitivity is key. The low performance is mostly believed to be due low representation of this class, with only around 40 cases for training. As expected, also from a clinical point-of-view, misclassifications were biased towards Sebhorreic Keratosis and Nevus.

V. Conclusion and Future Work

The last decades have witnessed significant progress in what concerns computer-aided diagnosis for several domains, namely dermatology and skin lesion classification. However, most breakthroughs are limited to well-controlled environments, with data acquired in very specific conditions, putting almost all effort in model development and improvements thereof. Despite all the impressive results found in the literature, most systems rely on non-standard acquisition equipment, handled by professionals whose focus is far from ensuring data quality and standardization.

To address this gap in real-world applications, this paper proposes different strategies for training a real-world image dataset for skin lesion classification, comprised of retrospective data from the Portuguese National Health System - DermAI. It presents 13 different differential diagnostic categories and its images are very diverse concerning the acquisition settings, the field-of-view, and overall quality, making it a challenging dataset, especially when compared to publicly available ones, as EDRA and Dermofit.

Different methodologies for training a Deep Neural Network on this unbalanced dataset were studied. First, we evaluated different network architectures with available pre-trained weights on ImageNet to assess with which one to proceed for the following experiments: MobileNet-v2, Resnet50 and EfficientNet-B3. Considering the trade-off between trainable parameters (model complexity) and performance (based on F1-score), we chose to proceed with the EfficientNet-B3.

Following, a model was trained from scratch, using only DermAI image samples. Unsurprisingly, these results were very poor (Accuracy of 14% and macro F1-score of 3%), with the model overfitting for two of the classes: Nevus and Haemangioma. A different, straightforward approach, relied on merging samples from common classes in the public skin lesion datasets into DermAI. In opposition, we also propose a sequential fine-tuning pipeline where the target dataset is fine-tuned after pre-training the model iteratively with other datasets, from larger and more general sets (ImageNet) to smaller, similar domain, ones, like EDRA and Dermofit. One of the first conclusions is that pre-training, even with datasets as general as ImageNet, have a significant impact on the model performance (Accuracy over 40% and macro F1-scores over 28%), even though the results reflect the complexity of the DermAI dataset.

When comparing the use of a merged dataset (using common skin lesion classes) with the use of sequential fine-tuning, the conclusions are less clear. In general, sequential fine-tuning returns marginally higher aggregated results (a marginal increase of macro F1-score), although some specific categories benefit from increased training samples in the same learning step. This is especially evident for classes where the images show significant differences between the datasets (e.g., Haemangioma or Pendulum Fibroma), since the first pre-training might lead the learning process away from extracting features that are more discriminative for categories only represented in the target dataset. On the other hand, other categories may benefit from such an iterative learning process, where features learned in other datasets help to generalize examples in the target dataset (e.g., Nevus and Malignant Melanoma). Moreover, although the averaged metrics favor a pre-training on EDRA and not including Dermofit, for some categories the latter shows to be beneficial, such as for Haemangioma, Neoplasm of Uncertain Behavior, and Viral Warts.

Our results highlight the challenges of a real-world application of skin lesion classification models, highly dependent on the available data, especially concerning its amount, quality, and diversity. One major conclusion is that high-quality, low-cost, and portable acquisition systems assume paramount importance in building good training sets. Furthermore, as disease cases are hard to get, in order to explore existing data, data-centric techniques should be explored in future work to improve results. These may include simply other aspect-ratio preserving resizing methods, or more interestingly, lesion segmentation methods to help the model focus on the most critical regions of the image without removing all of the, also important, surrounding context. Additionally, the use of neural activation maps or other related explainable methods could help analyse specific errors that may inspire further image preprocessing steps. Concerning data merging versus sequential fine-tuning, this work opens the door for future research on exploring both approaches simultaneously. It is possible to merge common classes (especially if that category's image variance is high) and follow with a sequential fine-tuning process for missing or not-shared classes. Further

 ISBN: 978-1-61208-916-4

future work considers including meta-data in the classification models and investigating joint training of multiple tasks related to skin lesion classification since parameter-sharing could be a good alternative to sequential pre-training. Finally, hierarchical classification can also be explored in future experiments, where different layers are considered towards improving the final classification. For example, one can first classify a sample into Benign, Malignant, or Uncertain categories and then into the corresponding final differential diagnosis based on the first decision.

ACKNOWLEDGMENT

This work was done under the scope of project "DERM.AI: Usage of Artificial Intelligence to Power Teledermatological Screening", and supported by national funds through 'FCT—Foundation for Science and Technology, I.P.', with reference DSAIPA/AI/0031/2018.

REFERENCES

[1] D.-G. for Health and F. Safety, "Market study on telemedicine," tech. rep., European Commission, 2018.

[2] S. Srinivas, R. K. Sarvadevabhatla, K. R. Mopuri, N. Prabhu, S. S. Kruthiventi, and R. V. Babu, "A taxonomy of deep convolutional neural nets for computer vision," *Frontiers in Robotics and AI*, vol. 2, p. 36, 2016.

[3] E. Valle, M. Fornaciali, A. Menegola, J. Tavares, F. V. Bittencourt, L. T. Li, and S. Avila, "Data, depth, and design: Learning reliable models for skin lesion analysis," *Neurocomputing*, vol. 383, pp. 303–313, 2020.

[4] C. N. Vasconcelos and B. N. Vasconcelos, "Convolutional neural network committees for melanoma classification with classical and expert knowledge based image transforms data augmentation," *arXiv preprint arXiv:1702.07025*, 2017. [retrieved: June,2021].

[5] L. T. Thao and N. H. Quang, "Automatic skin lesion analysis towards melanoma detection," in *2017 21st Asia Pacific symposium on intelligent and evolutionary systems (IES)*, pp. 106–111, IEEE, 2017.

[6] D. Moreira, P. Alves, F. Veiga, L. Rosado, and M. J. M. Vasconcelos, "Automated mobile image acquisition of macroscopic dermatological lesions," in *HEALTHINF*, pp. 122–132, 2021.

[7] J. Deng, W. Dong, R. Socher, L.-J. Li, K. Li, and L. Fei-Fei, "Imagenet: A large-scale hierarchical image database," in *2009 IEEE conference on computer vision and pattern recognition*, pp. 248–255, Ieee, 2009.

[8] V. Pomponiu, H. Nejati, and N.-M. Cheung, "Deepmole: Deep neural networks for skin mole lesion classification," in *2016 IEEE International Conference on Image Processing (ICIP)*, pp. 2623–2627, IEEE, 2016.

[9] K. M. Hosny, M. A. Kassem, and M. M. Foaud, "Classification of skin lesions using transfer learning and augmentation with alex-net," *PloS one*, vol. 14, no. 5, p. e0217293, 2019.

[10] A. R. Lopez, X. Giro-i Nieto, J. Burdick, and O. Marques, "Skin lesion classification from dermoscopic images using deep learning techniques," in *2017 13th IASTED international conference on biomedical engineering (BioMed)*, pp. 49–54, IEEE, 2017.

[11] K. Simonyan and A. Zisserman, "Very deep convolutional networks for large-scale image recognition," *arXiv preprint arXiv:1409.1556*, 2014. [retrieved: June,2021].

[12] D. Gutman, N. C. Codella, E. Celebi, B. Helba, M. Marchetti, N. Mishra, and A. Halpern, "Skin lesion analysis toward melanoma detection: A challenge at the international symposium on biomedical imaging (isbi) 2016, hosted by the international skin imaging collaboration (isic)," *arXiv preprint arXiv:1605.01397*, 2016. [retrieved: June,2021].

[13] A. Menegola, M. Fornaciali, R. Pires, F. V. Bittencourt, S. Avila, and E. Valle, "Knowledge transfer for melanoma screening with deep learning," in *2017 IEEE 14th International Symposium on Biomedical Imaging (ISBI 2017)*, pp. 297–300, IEEE, 2017.

[14] G. Argenziano, H. Soyer, V. De Giorgi, D. Piccolo, P. Carli, M. Delfino, and Others, "Dermoscopy: a tutorial," *EDRA, Medical Publishing & New Media*, 2002.

[15] K. Chatfield, K. Simonyan, A. Vedaldi, and A. Zisserman, "Return of the devil in the details: Delving deep into convolutional nets," in *Proceedings of the British Machine Vision Conference*, BMVA Press, 2014.

[16] EyePACs, "Diabetic retinopathy detection kagle." https://www.kaggle.com/c/diabetic-retinopathy-detection/data. [retrieved: June,2021].

[17] J. Kawahara, A. BenTaieb, and G. Hamarneh, "Deep features to classify skin lesions," in *2016 IEEE 13th international symposium on biomedical imaging (ISBI)*, pp. 1397–1400, IEEE, 2016.

[18] A. Krizhevsky, I. Sutskever, and G. E. Hinton, "Imagenet classification with deep convolutional neural networks," *Advances in Neural Information Processing Systems*, vol. 2, pp. 1097—1105, 2012.

[19] L. Ballerini, R. B. Fisher, B. Aldridge, and J. Rees, "A color and texture based hierarchical k-nn approach to the classification of non-melanoma skin lesions," in *Color Medical Image Analysis*, pp. 63–86, Springer, 2013.

[20] J. Kawahara, S. Daneshvar, G. Argenziano, and G. Hamarneh, "Seven-point checklist and skin lesion classification using multitask multimodal neural nets," *IEEE journal of biomedical and health informatics*, vol. 23, no. 2, pp. 538–546, 2018.

[21] C. Szegedy, V. Vanhoucke, S. Ioffe, J. Shlens, and Z. Wojna, "Rethinking the inception architecture for computer vision," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 2818–2826, 2016.

[22] T. Nedelcu, M. Vasconcelos, and A. Carreiro, "Multi-dataset training for skin lesion classification on multimodal and multitask deep learning," in *Proceedings of the 6th World Congress on Electrical Engineering and Computer Systems and Sciences (EECSS'20)*, pp. ICBES 120–1–8, 2020.

[23] N. C. Codella, D. Gutman, M. E. Celebi, B. Helba, M. A. Marchetti, S. W. Dusza, *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 *2018 IEEE 15th international symposium on biomedical imaging (ISBI 2018)*, pp. 168–172, IEEE, 2018.

[24] P. Tschandl, C. Rosendahl, and H. Kittler, "The ham10000 dataset, a large collection of multi-source dermatoscopic images of common pigmented skin lesions," *Scientific data*, vol. 5, no. 1, pp. 1–9, 2018.

[25] M. Combalia, N. C. Codella, V. Rotemberg, B. Helba, V. Vilaplana, O. Reiter, *et al.*, "Bcn20000: Dermoscopic lesions in the wild," *arXiv preprint arXiv:1908.02288*, 2019. [retrieved: June,2021].

[26] A. Mahbod, G. Schaefer, I. Ellinger, R. Ecker, A. Pitiot, and C. Wang, "Fusing fine-tuned deep features for skin lesion classification," *Computerized Medical Imaging and Graphics*, vol. 71, pp. 19–29, 2019.

[27] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 770–778, 2016.

[28] L. M. Abbott, R. Miller, M. Janda, H. Bennett, M. Taylor, C. Arnold, S. Shumack, H. P. Soyer, and L. J. Caffery, "Practice guidelines for teledermatology in australia," *Australasian Journal of Dermatology*, vol. 61, no. 3, pp. e293–e302, 2020.

[29] M. Sandler, A. Howard, M. Zhu, A. Zhmoginov, and L.-C. Chen, "Mobilenetv2: Inverted residuals and linear bottlenecks," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 4510–4520, 2018.

[30] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 770–778, 2016.

[31] M. Tan and Q. Le, "Efficientnet: Rethinking model scaling for convolutional neural networks," in *International Conference on Machine Learning*, pp. 6105–6114, PMLR, 2019.

[32] M. Lin, Q. Chen, and S. Yan, "Network in network," *arXiv preprint arXiv:1312.4400*, 2013. [retrieved: June,2021].

[33] T. Shermin, S. W. Teng, M. Murshed, G. Lu, F. Sohel, and M. Paul, "Enhanced transfer learning with imagenet trained classification layer," in *Pacific-Rim Symposium on Image and Video Technology*, pp. 142–155, Springer, 2019.

[34] A. I. Rubin, E. H. Chen, and D. Ratner, "Basal-cell carcinoma," *New England Journal of Medicine*, vol. 353, no. 21, pp. 2262–2269, 2005.

 ISBN: 978-1-61208-916-4