ORIGINAL ARTICLE

Improved diagnosis by automated macro- and micro-anatomical region mapping of skin photographs

L. Amruthalingam,[1,2] P. Gottfrois,[1] A. Gonzalez Jimenez,[1] B. Gökduman,[2] M. Kunz,[3] T. Koller,[4] DERMANATOMY Consortium,[3] M. Pouly,[4] A.A. Navarini,[3,*]

[1]Department of Biomedical Engineering, University of Basel, Basel, Switzerland
[2]Lucerne School of Computer Science and Information Technology, Lucerne University of Applied Sciences and Arts, Lucerne, Switzerland
[3]Department of Health Sciences and Technology, Swiss Federal Institute of Technology, Zurich, Switzerland
[4]Department of Dermatology, University Hospital of Basel, Basel, Switzerland
*Correspondence: A.A. Navarini. **E-mail:** alexander.navarini@usb.ch

Abstract

Background The exact location of skin lesions is key in clinical dermatology. On one hand, it supports differential diagnosis (DD) since most skin conditions have specific predilection sites. On the other hand, location matters for dermato-surgical interventions. In practice, lesion evaluation is not well standardized and anatomical descriptions vary or lack altogether. Automated determination of anatomical location could benefit both situations.

Objective Establish an automated method to determine anatomical regions in clinical patient pictures and evaluate the gain in DD performance of a deep learning model (DLM) when trained with lesion locations and images.

Methods Retrospective study based on three datasets: macro-anatomy for the main body regions with 6000 patient pictures partially labelled by a student, micro-anatomy for the ear region with 182 pictures labelled by a student and DD with 3347 pictures of 16 diseases determined by dermatologists in clinical settings. For each dataset, a DLM was trained and evaluated on an independent test set. The primary outcome measures were the precision and sensitivity with 95% CI. For DD, we compared the performance of a DLM trained with lesion pictures only with a DLM trained with both pictures and locations.

Results The average precision and sensitivity were 85% (CI 84–86), 84% (CI 83–85) for macro-anatomy, 81% (CI 80–83), 80% (CI 77–83) for micro-anatomy and 82% (CI 78–85), 81% (CI 77–84) for DD. We observed an improvement in DD performance of 6% (McNemar test *P*-value 0.0009) for both average precision and sensitivity when training with both lesion pictures and locations.

Conclusion Including location can be beneficial for DD DLM performance. The proposed method can generate body region maps from patient pictures and even reach surgery relevant anatomical precision, e.g. the ear region. Our method enables automated search of large clinical databases and make targeted anatomical image retrieval possible.

Received: 22 March 2022; Accepted: 30 June 2022

Conflicts of interest

None of the authors, including the DERMANATOMY Consortium, have conflicts of interest to declare.

Funding sources

Helmut-Fischer Foundation, Botnar Foundation, University of Basel.

Introduction

In clinical practice, the differential diagnosis (DD) of a skin lesion is influenced to a great extent by its anatomical location. Certain body regions are more likely than others to be affected by skin diseases, some of which have specific predilection sites.[1] Although this information is straightforward to obtain manually in clinical settings, it is more difficult to infer from patient pictures only, for example, in teledermatology context. The complexity increases the more zoomed-in the pictures and the less visible the anatomical landmarks are. An example of skin patches that are increasingly difficult to localize for human raters from image alone is shown in the Fig. S2. The ability to automatically localize small skin patches would also be useful for the automation of anatomical region mapping in skin photographs, as smaller skin patches are less likely to contain overlapping body parts, for example, folded arms over the trunk.

To be relevant in clinical settings, automated anatomical mappings should be more detailed than the main body regions and ideally reproduce the established international surface anatomy terminology.[2] Mohs micrographic surgery is a common operation in dermatology to remove cancerous lesions. In practice, surgeons are regularly confronted with situations where lesion's locations defined in a patient's profile are imprecise, sometimes wrong.[3] These mistakes happen due to the sheer number of different regions in the human anatomy and the difficulty of remembering them all, even for experienced clinicians. To avoid wrong site surgery, the anatomical description of biopsy sites is crucial as they may heal scar-free and the remaining tumour may become invisible.[4] Photographs might be unavailable to the surgeon, and patients may not be able to clarify biopsy sites, especially after several weeks delay for surgical appointment. With Mohs micrographic surgery, these issues are even more critical as it is a margin-controlled surgery, where there might not be a positive histological confirmation of the tumour right after the first stage of surgery. An automated system to assist clinicians with precise localization could benefit the documentation of biopsy locations.

Finally, another aspect to consider is the ever-increasing size of patient records and image databases kept for disease monitoring, future reference or research. The metadata of these images is often limited, restricting the usability of this data. To improve flexibility of these databases and accommodate new purpose of use, targeted image retrieval should be possible. Anatomical metadata would enable searching for specific regions of interest. However, producing such metadata manually is too costly in practice. With no automation in place, these valuable data sources remain underused.

Our study aimed to solve these challenges. We proposed a macro-anatomical deep learning model (DLM) to localize small skin patches on the main body regions, compared its performance with experts and showed that lesion location could improve classical DD DLM performance. Then, we trained a micro-anatomical DLM to segment the ear in its sub-regions, an approach that could assist dermatologists in lesion documentation. Both DLMs enable the generation at scale of the anatomical metadata required to perform targeted image retrieval.

Materials and methods

All images were obtained at the University Hospital of Zurich mainly from adult patients, type 1 to 3 on the Fitzpatrick scale. The data were anonymized by the removal of metadata and all personal identifying information. Subsequently, pictures were split and stored in small tiles (patches) precluding patient identification. Clinical images were taken at the same hospital with standard camera by a professional photographer. Capturing conditions were standardized: similar backgrounds and distances, controlled lighting and illumination. The visible anatomical region depended on lesions locations and were photographed mostly systematically. There were no artefacts such as pen markings, rulers or markings. We did not perform post- or pre- processing such as colour normalization, filtering or cropping (aside for the macro-anatomy location dataset).

Macro-anatomy

Body regions dataset The full dataset contained 6000 high-resolution patient's pictures showing the main body regions (Fig. S1): arms, legs, feet, hands, heads, and trunks. The initial training set, referred to as expert labelled (EL), contained 600 images (100 per body region) manually cropped to a single region. The remaining pictures composed the DL labelled (DLL) dataset. Their annotations were generated iteratively during the training process. We also included an "other" category of randomly selected pictures from the ImageNet[5] dataset to make the DLM robust against non-skin pictures such as clothes and background.

The images were cut into square patches with side length of 512 pixels corresponding to squares of 5–15 cm side length. This resulted in a training set composed of 277 122 DLL patches and 27 685 EL patches.

The DLM performance was evaluated on a separate test set of 140 independent images divided in 3570 strongly labelled patches. The body region distribution of the patches is available in the supplementary material. An example of a picture along with the DLM predictions is shown in the Fig. S3.

DLM training. The DLM was trained to localize each patch individually without having access to the rest of the image. We fine-tuned an EfficientNet[6] B2 DLM pre-trained on the ImageNet dataset with batch size 32 and input size 260 pixels for 40 epochs. We adopted a cyclic training approach inspired from Yalniz et al.[7] The DLM was first trained on EL patches with progressive resizing and used to predict the DLL set labels. Then, we retrained the DLM over the larger DLL dataset and fine-tuned with the EL patches. We repeated this cycle three times until the performance over the validation set stopped improving. During training, we scheduled the learning rate by applying the one cycle policy as suggested in Smith.[8]

Differential diagnosis from lesion image and macro-anatomical location

DD dataset We selected 16 skin diseases (detailed in Table 2) known to have specific predilection sites for a total of 3347 pictures. Diagnosis labels were provided by the photographer following dermatologists instructions who diagnosed patients in-person. The pictures repartition and usual predilection sites are presented in Table S5. The test set was generated by randomly sampling 20% of the pictures per disease ensuring no patient leak, which resulted in a total of 670 images.

DLM training We trained two DLMs based on the ResNet[9] architecture to perform the DD. Model A used only the lesion image, while Model B also had access to the lesion location predicted by the macro-anatomy DLM. To include this information, Model B learned a 128 dimensions embedding of the location, which was appended to the extracted lesion features (the following layer's size was adapted to account for this change). This is the only difference between both DLMs, which were trained following similar procedure, ImageNet pretraining, one cycle scheduling for the learning rate, with a batch size of 32, an input size of 512 pixels for 40 epochs.

Micro-anatomy

Ear segmentation dataset This dataset consisted of 182 ear photographs, each annotated for 12 different regions: anti-helix, anti-tragus, concha cavum, concha cymba, external auditory canal, helical root, helix, lobule, notch, scaphoid fossa, tragus, and triangular fossa. We also included the "non-ear" class to represent anything but ears. We kept 37 randomly selected pictures for the test set (ensuring no leak) to evaluate performance. An example of ear picture with its ground truth annotation is presented in Fig. 1.

DLM training We fine-tuned a U-Net[10] DLM with a ResNet backbone pre-trained on ImageNet. The training procedure was similar to the macro-anatomy DLM if we consider only the EL part of the cycle. The DLM was trained with an input size of 380 pixels, a batch size of 4 for 40 epochs.

Analysis

The performance of all DLMs was evaluated on the respective test sets using the average precision and sensitivity metrics (specificity available in the supplementary material) with 95% confidence interval determined using the non-parametric bootstrap resampling method.

In addition, for the macro-anatomy experiment, we randomly sampled 175 patches (25 per body region + the other category) from the test set, requested 6 dermatologists and 12 medical students to localize them and evaluated their performance similarly to the DLM.

For the DD experiment, we applied the McNemar's test to confirm whether the DLMs had significant difference in error proportions, following established practice for experiments with limited data.[11]

In the case of the micro-anatomy experiment, the average performance was evaluated on every pixel of the test images.

Results

Macro-anatomy

The DLM and experts performance are presented in Table 1, while Fig. 2 shows both confusion matrices. There was no

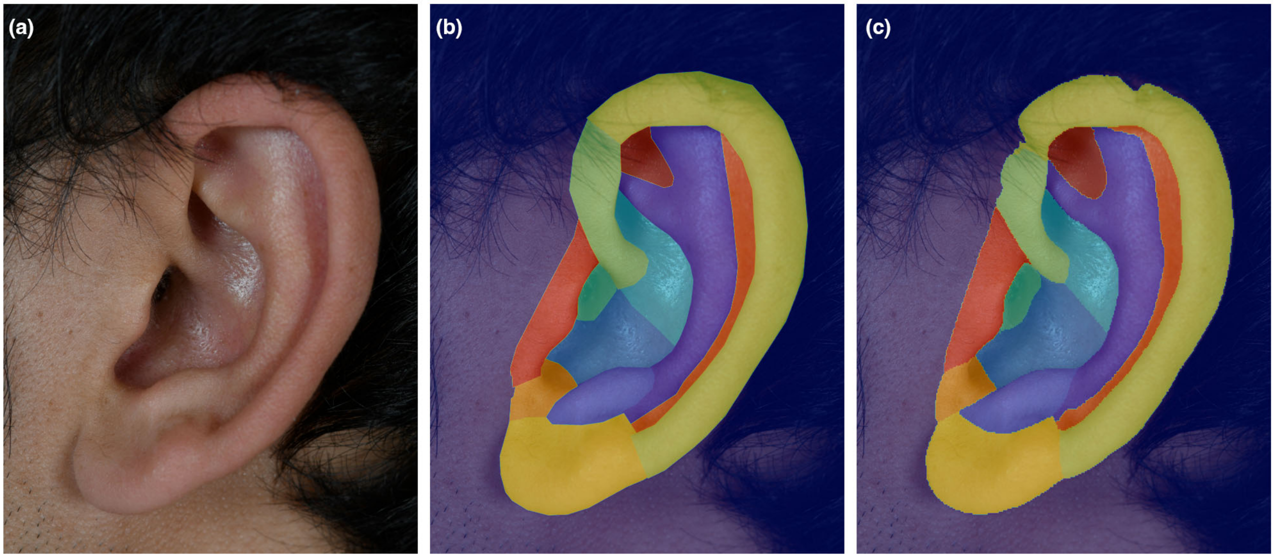


Figure 1 Ear test sample (a) with expert's annotations (b) and DLM's predictions (c). Picture randomly selected from the test set. The original image is shown in (a), the expert's annotation in (b) and the DLM's predictions in (c). The regions are coloured as follows: anti-helix in violet, anti-tragus in light violet, concha cavum in blue, concha cymba in light blue, external auditory canal in green, helical root in light green, helix in light yellow, lobule in yellow, notch in light orange, scaphoid fossa in orange, tragus in red, triangular fossa in light brown, non-ear in dark shade.

Table 1 Macro-anatomy performance

Region	DLM			Experts		
	Test images	Precision	Sensitivity	Test images	Precision	Sensitivity
Arm	510	75% (72–80)	77% (74–80)	25	44% (24–83)	35% (13–54)
Leg	510	80% (75–84)	69% (65–72)	25	49% (34–65)	42% (26–57)
Feet	510	86% (83–89)	88% (86–91)	25	78% (50–97)	50% (31–66)
Hand	510	93% (90–95)	84% (80–87)	25	62% (44–82)	71% (49–90)
Head	510	89% (86–92)	94% (92–96)	25	68% (42–90)	48% (28–77)
Other	510	100% (100–100)	99% (98–99)	25	91% (79–100)	100% (100–100)
Trunk	510	70% (66–74)	80% (77–84)	25	39% (27–59)	55% (22–81)
Average	-	85% (84–86)	84% (83–85)	–	62% (56–70)	57% (52–65)

Performance evaluated on the full test set for the DLM and on a stratified random sample of the test set for the expert panel composed of 6 dermatologists and 12 students. The values in parentheses are the 95% confidence interval. For the experts, the performance reported is the average of all individual performances.

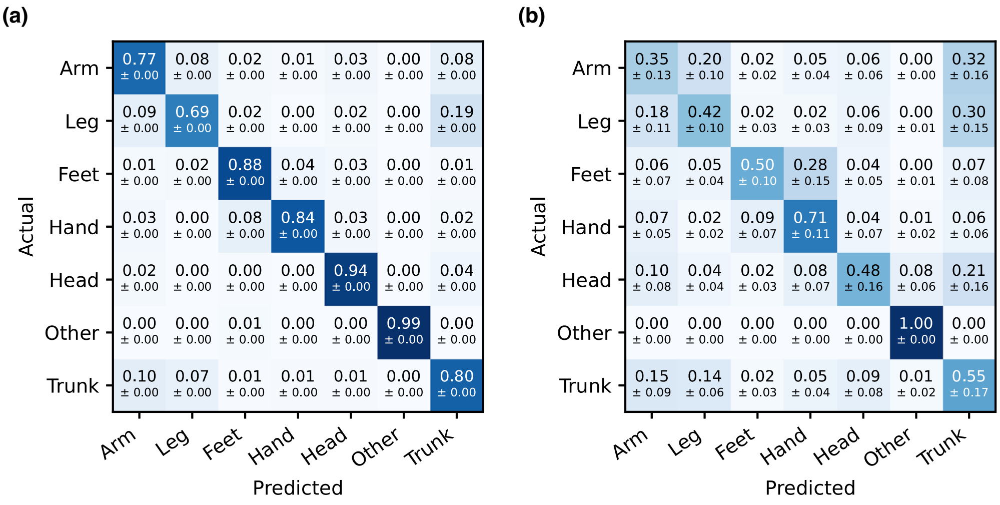


Figure 2 Confusion matrices for the macro-anatomy DLM (a) and the experts (b). The values show the proportion of patches ± SD. The average proportion ± SD of the patches localized among the six body regions and the "other" class. The vertical axis shows the true labels of the patches while the horizontal axis shows the predicted labels. The diagonal values correspond to the sensitivity for the body regions.

significant difference between the performance of dermatologists and medical students (Table S2).

The DLM reached an average precision of 85% (CI 84–86) and an average sensitivity of 84% (CI 83–85). In contrast, the average of experts' precision was 62% (CI 56–70) and for sensitivity 57% (CI 52–65).

Unsurprisingly, the DLM could almost flawlessly differentiate skin picture from non-skin pictures. The different body regions were well discriminated by the DLM, the best example being the patches coming from the head region, which were rarely confused (~6%) with any other classes. Leg was the worst performing class, confused with either arms or trunk and *vice versa*.

The experts' large standard deviation (Fig. 2b) for each region indicates an important inter-individual variation and thus highlights the lack of consensus. The confusion matrix shows difficulties with the trunk, arm and leg regions. The relatively higher sensitivity of the trunk region and its lower precision when compared with the legs and arms indicates that participants tended to default to the trunk region when no clear cues were available. The confusion of the trunk with the head region was due to patches showing skin from the cheeks. Feet were also mistaken with hands, but the opposite occurred less frequently. Two to three patches from the head containing mainly hairs were mistaken with the non-skin class.

Table 2 Differential diagnosis performance

Disease	Test images	Precision A	Sensitivity A	Precision B	Sensitivity B
Acne	48	84% (74–94)	77% (66–88)	88% (74–96)	73% (63–83)
Drug eruptions	43	85% (74–94)	79% (66–89)	97% (93–100)	86% (73–95)
Darier disease	14	64% (33–89)	50% (24–72)	67% (33–91)	57% (32–83)
Dyshidrotic eczema	50	77% (66–88)	88% (80–96)	87% (78–96)	94% (88–100)
Nummular dermatitis	34	79% (68–90)	88% (75–97)	84% (72–93)	91% (78–100)
Hand eczema	50	74% (63–84)	74% (62–85)	76% (66–86)	82% (70–92)
Impetigo	19	76% (56–97)	68% (44–92)	88% (71–100)	79% (55–98)
Melasma	42	60% (44–74)	67% (51–80)	57% (41–69)	74% (58–89)
Morphea	68	84% (75–91)	75% (66–84)	97% (91–100)	88% (81–96)
Onychomycosis	60	81% (72–91)	90% (83–97)	85% (79–94)	88% (81–96)
Palmoplantar keratoderma	45	85% (73–93)	73% (60–84)	92% (83–98)	76% (64–85)
Pityriasis rosea	50	74% (62–83)	84% (75–92)	78% (67–88)	94% (84–99)
Rosacea	49	79% (67–91)	69% (55–83)	75% (63–87)	73% (63–84)
Tinea pedis	27	71% (49–89)	56% (40–74)	84% (71–98)	78% (63–94)
Ulcer	41	90% (81–99)	93% (79–100)	95% (87–100)	93% (79–100)
Vitiligo	40	61% (49–72)	70% (58–82)	62% (49–76)	62% (47–75)
Average	–	76% (73–80)	75% (72–79)	82% (78–85)	81% (77–84)

Performance evaluated on a 20% random sample of the images for each diagnosis (ensuring no patient leak). Model A was trained with lesion pictures only, while model B also had access to the lesions' locations. The 95% confidence interval is shown in parentheses.

Differential diagnosis from lesion image and macro-anatomical location

The performance of both DLMs is presented in Table 2. Model B reached an average precision and sensitivity of 82% (CI 78–85) and 81% (CI 77–84). Compared with model A, which achieved 76% (CI 73–80) and 75% (CI 72–79) for average precision and sensitivity, this represents an average improvement of 6% for both metrics.

The McNemar's test applied to the full test set confirmed that both classifiers had significant difference in error proportions with *P*-value 0.0009.

We observed a reduction of the sensitivity for acne, onychomycosis and vitiligo in model B. This was due to confusions with diseases sharing similar predilection sites (see confusion matrices in Figs. S4-S5), for example, the head for acne with rosacea, melasma and impetigo. The drop in precision for melasma and rosacea can be explained similarly.

Micro-anatomy

The performance of the ear segmentation DLM is presented in Fig. 3.

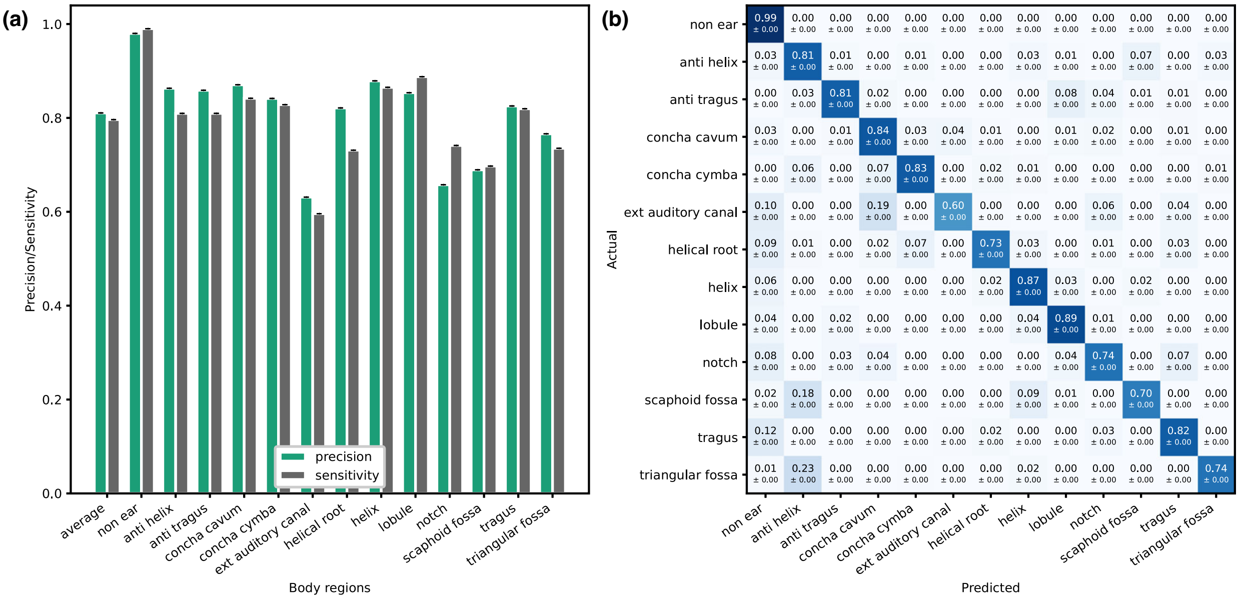
The DLM reached an average precision of 81% (CI 80–83) and an average sensitivity of 80% (CI 77–83). The most challenging classes were the external auditory canal, notch and scaphoid fossa. These were also the smallest regions with less training samples in comparison to the other classes. Depending on the ear type and orientation, they could be absent or very small in comparison with neighbouring regions.

Discussion

We addressed the challenge to automatically map skin pictures to their corresponding anatomical regions. A macro-anatomy DLM was trained using a dataset of 6′000 patient images to map small skin patches to the corresponding body regions. An expert panel of 18 dermatologists and medical students performed a similar task with lower precision and sensitivity and with high inter-rater variability. We showed that lesions location could improve DD DLM performance. Finally, we presented a micro-anatomy DLM able to segment ear pictures precisely enough for surgery applications.

Previous studies on anatomy segmentation with DL have focused on 3D CT scans to identify body parts and organs.[12,13] While there have been studies on geographical mapping of photographs' origin using DL on a global scale,[14] our study is to the best of our knowledge the first attempt to do the same on the human body surface from standard photographs. The combined use of lesion location and image for DD were limited so far to skin cancer studies,[15,16] which also leveraged other patient clinical features such as age and gender, yielding improved DD accuracy. Lesion location was also used as secondary objective in multi-task learning context to improve performance of lesion morphology classification.[17]

One design limitation of this study is to restrict the DD experiment to diseases with specific predilection sites. In future work we will confirm if the reported performance improvement also holds when including other diagnoses without this constraint. This study is also limited by its choice of macro-anatomy body



	non ear	anti helix	anti tragus	concha cavum	concha cymba	ext auditory canal	helical root	helix	lobule	notch	scaphoid fossa	tragus	triangular fossa
non ear	0.99 ± 0.00	0.00 ± 0.00	0.00 ± 0.00	0.00 ± 0.00	0.00 ± 0.00	0.00 ± 0.00	0.00 ± 0.00	0.00 ± 0.00	0.00 ± 0.00	0.00 ± 0.00	0.00 ± 0.00	0.00 ± 0.00	0.00 ± 0.00
anti helix	0.03 ± 0.00	0.81 ± 0.00	0.01 ± 0.00	0.00 ± 0.00	0.01 ± 0.00	0.00 ± 0.00	0.00 ± 0.00	0.03 ± 0.00	0.01 ± 0.00	0.00 ± 0.00	0.07 ± 0.00	0.00 ± 0.00	0.03 ± 0.00
anti tragus	0.00 ± 0.00	0.03 ± 0.00	0.81 ± 0.00	0.02 ± 0.00	0.00 ± 0.00	0.00 ± 0.00	0.00 ± 0.00	0.00 ± 0.00	0.08 ± 0.00	0.04 ± 0.00	0.01 ± 0.00	0.01 ± 0.00	0.00 ± 0.00
concha cavum	0.03 ± 0.00	0.00 ± 0.00	0.01 ± 0.00	0.84 ± 0.00	0.03 ± 0.00	0.04 ± 0.00	0.01 ± 0.00	0.00 ± 0.00	0.01 ± 0.00	0.02 ± 0.00	0.00 ± 0.00	0.01 ± 0.00	0.00 ± 0.00
concha cymba	0.00 ± 0.00	0.06 ± 0.00	0.00 ± 0.00	0.07 ± 0.00	0.83 ± 0.00	0.00 ± 0.00	0.02 ± 0.00	0.01 ± 0.00	0.00 ± 0.00	0.00 ± 0.00	0.00 ± 0.00	0.00 ± 0.00	0.01 ± 0.00
ext auditory canal	0.10 ± 0.00	0.00 ± 0.00	0.00 ± 0.00	0.19 ± 0.00	0.00 ± 0.00	0.60 ± 0.00	0.00 ± 0.00	0.00 ± 0.00	0.00 ± 0.00	0.06 ± 0.00	0.00 ± 0.00	0.04 ± 0.00	0.00 ± 0.00
helical root	0.09 ± 0.00	0.01 ± 0.00	0.00 ± 0.00	0.02 ± 0.00	0.07 ± 0.00	0.00 ± 0.00	0.73 ± 0.00	0.03 ± 0.00	0.00 ± 0.00	0.01 ± 0.00	0.00 ± 0.00	0.03 ± 0.00	0.00 ± 0.00
helix	0.06 ± 0.00	0.00 ± 0.00	0.00 ± 0.00	0.00 ± 0.00	0.00 ± 0.00	0.00 ± 0.00	0.02 ± 0.00	0.87 ± 0.00	0.03 ± 0.00	0.00 ± 0.00	0.02 ± 0.00	0.00 ± 0.00	0.00 ± 0.00
lobule	0.04 ± 0.00	0.00 ± 0.00	0.02 ± 0.00	0.00 ± 0.00	0.00 ± 0.00	0.00 ± 0.00	0.00 ± 0.00	0.04 ± 0.00	0.89 ± 0.00	0.01 ± 0.00	0.00 ± 0.00	0.00 ± 0.00	0.00 ± 0.00
notch	0.08 ± 0.00	0.00 ± 0.00	0.03 ± 0.00	0.04 ± 0.00	0.00 ± 0.00	0.00 ± 0.00	0.00 ± 0.00	0.00 ± 0.00	0.04 ± 0.00	0.74 ± 0.00	0.00 ± 0.00	0.07 ± 0.00	0.00 ± 0.00
scaphoid fossa	0.02 ± 0.00	0.18 ± 0.00	0.00 ± 0.00	0.00 ± 0.00	0.00 ± 0.00	0.00 ± 0.00	0.00 ± 0.00	0.09 ± 0.00	0.01 ± 0.00	0.00 ± 0.00	0.70 ± 0.00	0.00 ± 0.00	0.00 ± 0.00
tragus	0.12 ± 0.00	0.00 ± 0.00	0.00 ± 0.00	0.00 ± 0.00	0.00 ± 0.00	0.00 ± 0.00	0.02 ± 0.00	0.00 ± 0.00	0.00 ± 0.00	0.03 ± 0.00	0.00 ± 0.00	0.82 ± 0.00	0.00 ± 0.00
triangular fossa	0.01 ± 0.00	0.23 ± 0.00	0.00 ± 0.00	0.00 ± 0.00	0.00 ± 0.00	0.00 ± 0.00	0.00 ± 0.00	0.02 ± 0.00	0.00 ± 0.00	0.00 ± 0.00	0.00 ± 0.00	0.00 ± 0.00	0.74 ± 0.00

Figure 3 Ear segmentation DLM's micro-anatomical performance. The values show the proportion of images ± SD. (a) Precision and sensitivity: the average pixel precision and sensitivity reached on the test set by the DLM. (b) Confusion matrix: the average pixel proportion segmented among the 12 ear regions and the "non-other" class achieved by the DLM. The vertical axis shows the true pixel labels while the horizontal axis shows the pixel labels predicted by the DLM. The diagonal values correspond to the sensitivity for the ear regions.

regions, which is not sufficiently precise for dermatological description of lesions. The natural improvement is to refine the taxonomy. Approaches similar to the proposed micro-anatomy DLM for ears can be applied to other regions, which we plan to do in future work as well. Finally, another limitation of this study comes from the standardized nature of the data used to train the DLMs. All training images came from the same hospital and were taken with similar lighting, zoom and patient posture.

Following the CLEAR guidelines,[18] we determined the following bias sources in our study. There was a relative class imbalance between some of the diagnoses, which we mainly mitigated during dataset preparation by capping the total number of images per diagnoses (images were selected randomly). We chose not to vary the class distribution between the train and test set due to the limited amount of available pictures. The achieved performance showed that the minority classes (Darier disease, Impetigo and Tinea pedis) were not overlooked by the DLM and did benefit from the addition of lesions location.

Patients included in our datasets mainly had skin type 1 to 3 on the Fitzpatrick scale, implying that our DLM performance are valid only on patients with this skin pigmentation. Unfortunately no patient-level image metadata was available, which precluded the evaluation of related biases and constitutes a theoretical limit of this study.

Finally, since the chosen diseases had specific predilection sites, the images showed different anatomical parts, e.g., acne pictures always included patients heads, causing a bias. This was mitigated by selecting skin diseases such that each of the main body regions were among the predilection sites of at least four different diseases.

In direct application of our study, we generated both the macro- and micro-anatomical metadata of our institutes dermatology database (over 180 000 images), fully automatically and with no time-consuming manual intervention, illustrating the scalability and applicability of our approach. While the whole analysis was performed in <6 h with our DLMs, we estimate that one human annotator would require a minimum of 763 working days for the macroanatomical mapping (2 min per images) and 32 days for the microanatomical mapping of the ear pictures (10 min per images). With this metadata, the dermatology institute can now query its database for full or cropped pictures containing specific body regions or ear sub-regions. Since diagnosis is usually kept as metadata, a practical example of image retrieval would be to look for cases of eczema located on the leg: a first filter would return the available images diagnosed with eczema, followed by a second filter, which would extract the leg region.

An error analysis revealed that the DLMs performance were lower when images were captured in too dissimilar conditions or from specific regions (genitals, tongue, *etc.*). This drawback is faced by all deep learning (DL) approaches and can be tackled by fine-tuning the DLM on an external validation set acquired under the same conditions. This process would directly start with the DLMs' parameters learned in this study instead of the ones obtained on ImageNet, effectively reducing training costs and dataset size requirements.

While lesions locations could theoretically also be extracted by text mining patients records, this information should be accurately documented and properly linked to the corresponding patients images, which is not usually the case in clinical practice where reported locations can be imprecise.[3,4] One of our DLMs purposes is especially to assist clinicians in reporting accurate locations. The DLMs presented in this work can be regarded as a building block for future automated DD systems. One open issue with current photo diagnosis systems is that by fully relying on the capacities of DLMs to autonomously find features and learn how to combine them, researchers are not able to understand the algorithms' decision process anymore as the complexity of the DLMs grow. An alternative would be to base DL systems on the actual DD processes (usually decision trees) followed by dermatologists and use different DLMs for each step in the decision tree. Clinicians could then inspect and validate the intermediate DLMs' predictions to better understand the final recommendation of the system. As with any differential diagnosis, this starts with the location on the body.

Acknowledgement

The patients recruited for this study have given written informed consent to the publication of their case details. Open access funding provided by Universitat Basel. Open access funding provided by Universitat Basel.

Ethical approval

Swiss ethical permission (EKNZ, 2018–01074).

Disclosure statement

A. A. Navarini declares being a consultant and advisor and/or receiving speaking fees and/or grants and/or served as an investigator in clinical trials for AbbVie, Almirall, Amgen, Biomed, BMS, Boehringer Ingelheim, Celgene, Eli Lilly, Galderma, GSK, LEO Pharma, Janssen-Cilag, MSD, Novartis, Pfizer, Pierre Fabre Pharma, Regeneron, Sandoz, Sanofi, and UCB.

Author contributions

Ludovic Amruthalingam: Conceptualization, Data curation, Formal analysis, Investigation, Methodology, Project administration, Software, Validation, Visualization, Writing – original draft, Writing – review & editing. Philippe Gottfrois: Methodology, Validation, Writing – review & editing. Alvaro Gonzalez Jimenez: Methodology, Validation, Writing – review & editing. Bulus Gökduman: Data curation. Michael Kunz: Data curation, Methodology, Validation, Writing – review & editing. Thomas Koller: Methodology, Resources, Supervision, Validation, Writing – review & editing. DERMANATOMY Consortium: Data curation. Marc Pouly: Conceptualization, Funding acquisition, Methodology, Project administration, Resources, Supervision, Validation, Writing – review & editing. Alexander A. Navarini: Conceptualization, Data curation, Funding acquisition, Methodology, Project administration, Resources, Supervision, Validation, Writing – review & editing.

Consortia

DERMANATOMY Consortium: Julia-Tatjana Maul; Lara V. Maul; Lisa Kostner; Dagmar Jamiolkowski; Barbara Erni; Christophe Hsu; Nina Meienberger; M. Nicolas Khouri; M. Christiane Palm; M. Damian Wuethrich; Madeleine Anliker; M. Manabu Rohr; Matija Horvat; Noemie Eckert; M. Kei Mathis; M. Salvatore Conticello; Sijamini Baskaralingam; Lea Rotondi; M. Pascal Kobel.

Data availability statement

Under Swiss regulations, this study's ethical permission (2018, 01074, EKNZ) did not include sharing patients images.

References

1 Ruocco V, Ruocco E, Brunetti G, Sangiuliano S, Wolf R. Opportunistic localization of skin lesions on vulnerable areas. *Clin Dermatol* 2011; **29**: 483–488.

2 Kenneweg KA, Halpern AC, Chalmers RJ, Soyer HP, Weichenthal M, Molenda MA. Developing an international standard for the classification of surface anatomic location for use in clinical practice and epidemiologic research. *J Am Acad Dermatol* 2019; **80**: 1564–1584.

3 Ochoa SA, Lawrence N. Availability of biopsy site documentation for Mohs surgery. *J Dermatol Nurses Assoc* 2015; **7**: 273–276.

4 Zhang J, Rosen A, Orenstein L *et al*. Factors associated with biopsy site identification, postponement of surgery, and patient confidence in a dermatologic surgery practice. *J Am Acad Dermatol* 2016; **74**: 1185–1193.

5 Deng J, Dong W, Socher R, Li LJ, Li K, Fei-Fei L. Imagenet: a large-scale hierarchical image database. In Grauman K, eds. 2009 IEEE Conference on Computer Vision and Pattern Recognition, IEEE Computer Society, Miami, FL, 2009: 248–255.

6 Tan M, Le Q. Efficientnet: rethinking model scaling for convolutional neural networks. In Lawrence N, eds. International Conference on Machine Learning, Proceedings of Machine Learning Research (PLMR), Long Beach, CA, 2019: 6105–6114.

7 Yalniz IZ, Jégou H, Chen K, Paluri M, Mahajan D. *Billion-scale semi-supervised learning for image classification. arXiv preprint arXiv:1905.00546*. 2019.

8 Smith LN. *A disciplined approach to neural network hyper-parameters: part 1-learning rate, batch size, momentum, and weight decay. arXiv preprint arXiv:1803.09820*. 2018.

9 He K, Zhang X, Ren S, Sun J. Deep residual learning for image recognition. In Grauman K, eds. Proceedings of the IEEE conference on computer vision and pattern recognition, IEEE Computer Society Las Vegas, NV, 2016: 770–778.

10 Ronneberger O, Fischer P, Brox T. U-net: Convolutional networks for biomedical image segmentation.In Navab N, Hornegger J, eds. International Conference on Medical Image Computing and Computer-Assisted Intervention, Springer-Verlag, Munich, 2015: 234–241.

11 Dietterich TG. Approximate statistical tests for comparing supervised classification learning algorithms. *Neural Comput* 1998; **10**: 1895–1923.

12 Zhu W, Huang Y, Zeng L *et al*. AnatomyNet: deep learning for fast and fully automated whole-volume segmentation of head and neck anatomy. *Med Phys* 2019; **46**: 576–589.

13 Liu L, Wolterink JM, Brune C, Veldhuis RN. Anatomy-aided deep learning for medical image segmentation: a review. *Phys Med Biol* 2021; **66**: 11.

14 Weyand T, Kostrikov I, Philbin J Planet-photo geolocation with convolutional neural networks. In Leibe B, eds. European Conference on Computer Vision, Springer, Amsterdam, 2016 37–55.

15 Nunnari F, Bhuvaneshwara C, Ezema AO, Sonntag D. A study on the fusion of pixels and patient metadata in CNN-based classification of skin lesion images. In Holzinger A, eds. International Cross-Domain Conference for Machine Learning and Knowledge Extraction, Springer, Cham, 2020: 191–208.

16 Pacheco AG, Krohling RA. An attention-based mechanism to combine images and metadata in deep learning models applied to skin cancer classification. *IEEE J Biomed Health Inform* 2021; **25**: 3554–3563.

17 Liao H, Luo J. *A deep multi-task learning approach to skin lesion classification. arXiv preprint arXiv:1812.03527*. 2018.

18 Daneshjou R, Barata C, Betz-Stablein B *et al*. Checklist for evaluation of image-based artificial intelligence reports in dermatology: CLEAR derm consensus guidelines from the international skin imaging collaboration artificial intelligence working group. *JAMA Dermatol* 2022; **158**: 90–96.

Supporting information

Additional Supporting Information may be found in the online version of this article:

Figure S1. Body regions.

Figure S2. Examples of skin patches with increasing localization difficulty.

Figure S3. Generation of macro-anatomical body regions mapping.

Figure S4. Confusion matrix of DLM trained with lesion's pictures only. The values show the proportion of images ± SD.

Figure S5. Confusion matrix of DLM trained with both lesion's locations and pictures. The values show the proportion of images ± SD.

Table S1. Body regions dataset patch distribution.

Table S2. Macro-anatomy experts performance.

Table S3. Macro-anatomy DLM performance using only strongly labeled data.

Table S4. Differential diagnosis specificity.

Table S5. Predilection sites of the different diagnoses.