Deep Learning in Clinical Dermatology

Inaugural dissertation

to
be awarded the degree of

Dr. sc. med.

presented at the Faculty of Medicine of the University of Basel

by
Ludovic Hadrien Amruthalingam
from Plan-les-Ouates, Canton of Geneva

Basel, 2023

Approved by the Faculty of Medicine on application of
Prof. Dr. Alexander A. Navarini, University of Basel – *primary advisor*
Prof. Dr. Marc Pouly, Lucerne University of Applied Sciences and Arts – *secondary advisor*
Prof. Dr. Philipp Tschandl, Medical University of Vienna – *external expert*
Prof. Dr. Thomas Koller, Lucerne University of Applied Sciences and Arts – *advisor*
Prof. Dr. Philippe C. Cattin, University of Basel – *defense chair*

Basel, the 28. Juni 2022

Prof. Dr. Primo Schär – *Dean*

Contents

Acronyms

AI	artificial intelligence.
ANN	artificial neural network.
CNN	convolutional neural network.
DL	deep learning.
DLM	deep learning model.
GAN	generative adversarial network.
GDPR	general data protection regulation.
ISIC	international skin imaging collaboration.
IWC	ichthyosis with confetti.
ML	machine learning.
PASI	psoriasis area and severity index.
PPP	palmoplantar pustular psoriasis.

Acknowledgements

I would like to thank Alexander Navarini and Marc Pouly for their precious supervision and for providing a great work environment throughout this thesis. I also thank Thomas Koller for his expert advising in machine vision, Philipp Tschandl for reviewing the full thesis and Philippe Cattin for chairing the defense. I want to thank all my colleagues from the Digital Dermatology group, Navarinilab, ABIZ group, DBE and USB for their ideas, suggestions, reviews, and recommendations. I also thank our interns and students who made our projects possible thanks to their hard work. Finally, I want to thank my family for their advice and precious support.

Summary / Zusammenfassung

Summary

The prevalence of skin diseases is high. A recent survey reported that half of the European population was afflicted with skin conditions. However, the resulting demand for dermatological care cannot be met satisfactorily because of a general shortage of dermatologists that will realistically not be filled by the healthcare sector. Alternative solutions should therefore be pursued to increase the capacities of the current healthcare workforce.

The recent progress of machine vision enabled by deep learning has allowed researchers to automate parts of dermatologists' workflow with an effective scale-up potential. In this work, we present different approaches based on deep learning that either include aspects of dermatologists' workflow or whose predictions can easily be verified by clinicians. We propose a method for the generation of anatomical maps from patient photographs to assist dermatologists with lesion documentation and enable lesion detection and segmentation systems to stratify their predictions anatomically. Based on key features from lesion dermatological description, we develop an approach for the differential diagnosis of skin diseases. To enable objective severity assessment, we propose a method for the segmentation and quantification of palmoplantar pustular psoriasis, ichthyosis with confetti and hand eczema. Combined with the anatomy approach, we generate the anatomical stratification of hand eczema lesions. To concretize our research efforts, we present an African teledermatology initiative aiming to provide semi-automatic triage of the six most prevalent local skin diseases. Finally, we introduce our framework to enable researchers with medical background to train and evaluate deep learning models.

Zusammenfassung

Die Prävalenz von Hautkrankheiten ist hoch. Einer kürzlich durchgeführten Umfrage zufolge leidet die Hälfte der europäischen Bevölkerung an Hautkrankheiten. Die daraus resultierende Nachfrage nach dermatologischer Versorgung kann jedoch nicht erfüllt werden, da ein allgemeiner Mangel an Dermatologinnen und Dermatologen besteht, der realistischerweise nicht durch den Gesundheitssektor ausgeglichen werden kann. Daher sollten alternative Lösungen angestrebt werden, um die Kapazitäten des derzeitigen Gesundheitspersonals zu vervielfachen.

Die jüngsten Fortschritte im Bereich des maschinellen Sehens dank Deep Learning haben es ermöglicht, Teile der Arbeitsabläufe von Dermatologinnen und Dermatologen zu automatisieren, und zwar mit einem effektiven Skalierungspotenzial. In dieser Arbeit stellen wir verschiedene auf Deep Learning basierende Ansätze vor, die entweder Aspekte des Arbeitsablaufs von Dermatologinnen und Dermatologen einbeziehen oder deren Vorhersagen leicht überprüft werden können. Wir schlagen eine Methode zur Erzeugung anatomischer Karten aus Patientenfotos vor, um die Dokumentation von Läsionen zu unterstützen und es Systemen zur Läsionserkennung und -segmentierung zu ermöglichen, ihre Vorhersagen anatomisch zuzuordnen. Auf der Grundlage von Schlüsselmerkmalen aus der dermatologischen Beschreibung von Läsionen entwickeln wir einen Ansatz für die Differentialdiagnose von Hautkrankheiten. Um eine objektive Bewertung des Schweregrads zu ermöglichen, schlagen wir eine Methode zur Segmentierung und Quantifizierung von palmoplantarer pustulöser Psoriasis, Ichthyose mit Konfetti und Handekzemen vor. In Kombination mit dem anatomischen Ansatz implementieren wir die feingranulare anatomische Zuordnung von Handekzemläsionen. Um unsere Forschungsbemühungen zu konkretisieren, stellen wir eine afrikanische Teledermatologie-Initiative vor, die eine halbautomatische Triage der sechs häufigsten lokalen Hautkrankheiten ermöglichen soll. Schließlich stellen wir unser Framework vor, das es Forscherinnen und Forschern mit medizinischem Hintergrund ermöglicht, Deep-Learning-Modelle zu trainieren und zu bewerten.

Chapter 1

Introduction

Motivation

The skin is the largest organ of the human body, serving as barrier against pathogens and physical harm. There are over two thousand different skin conditions, some of which are highly prevalent. Half of the adult European population [50] suffers from skin diseases, resulting in a high demand for dermatological care. However, it takes twelve years of education to train dermatologists to diagnose skin diseases and treat them. These tasks are time-intensive. They are performed without decision support, and quality depends on the practitioner's skills and experience. Proficiency comes with practice and the experience of thousands of patient cases. Currently, there is a general shortage of dermatologists [97] even in high-income countries, where patients can face important delays [161].

The skin is easily accessible and diseases can be photographed with standard cameras. Since the analysis of skin diseases is based on visual inspection, there is an increasing adoption of teledermatology as an effective mean to scale-up patient care and serve remote geographical regions [102, 149, 191]. However, the number of attended patients remains bounded by the limited number of dermatologists, who need to process each patient case individually. The recent progress of machine vision, especially with deep learning, has paved the way for automation, which seems today the most promising solution to scale-up dermatological care in general [9]. Based on available dermatology databases, deep learning models can be trained to automatically perform repetitive tasks such as triage, severity grading and differential diagnosis of simple patient cases. Deep learning synergizes well with teledermatology and has the potential to expand patient coverage with the same workforce. Automation also improves the quality of care, as it reduces variability between dermatologists and increases precision and reproducibility.

Contribution

In this thesis, we present deep learning methods to automate and improve parts of dermatologists' workflow:

1. **Anatomy mapping:** We propose a method to automatically map anatomical regions in patient photographs at different levels of precision, starting from the main body regions to surgery-relevant anatomical description. This system can assist in medical education, support dermatologists with lesion documentation, and synergize with lesion detection and segmentation applications to produce anatomical stratification. In addition, we can generate the anatomical metadata of dermatology databases and enable targeted image retrieval based on specific anatomical regions.

2. **Differential diagnosis:** Inspired by dermatologists' decision processes, we present a hybrid approach based on lesion images and features from their dermatological description to automate differential diagnosis. We show that training deep learning models with both lesions' locations and images can improve their diagnosis performance. Similarly, we show that leveraging lesions' morphology also has a positive effect on performance. Since both lesion location and morphology can be verified by dermatologists, our approach enables them to validate or correct these features, thereby gaining some control and understanding of the models' workflow.

3. **Severity grading:** Aiming for automated disease severity assessment, we present a method to objectively quantify the lesions of pustular diseases with a specific focus on palmoplantar pustular psoriasis. Our system generates pustule counts and surface estimates, both of which are surrogate markers of disease severity. Furthermore, we apply this method to automatically segment lesions of ichthyosis with confetti and hand eczema. In synergy with our anatomy system, we partially automate the hand eczema severity index and present novel results on the disease anatomical distribution.

4. **Synthetic data generation:** To solve the general problem of dermatology data availability, we explore the capacities of generative adversarial networks to generate synthetic clinical and dermoscopy images of skin lesions.

5. **Teledermatology:** We describe our research teledermatology initiative in sub-Saharan countries, where we aim to deploy automated triage solutions for the most frequent skin conditions. All collected data will be anonymized and made available to the research community.

6. **Model development framework:** Finally, we give an overview of our training and evaluation framework, which enables researchers with medical background to participate in the development of deep learning models.

The methods described in this work are all based on patient's photographs. They were developed under the constraint to only use models with relatively low computational requirements that could realistically be deployed in clinical practice and maintained over time. We intentionally avoided very deep and therefore resource intensive neural networks as well as complex ensemble models for ease of industrialization. Our research did not include the use of text data, since we aimed to develop methods that could be used in clinical or teledermatology consultations without prior knowledge about the patient. Text data was also not readily available and would have needed a specific ethical permission.

Outline

We start by introducing both the fields of clinical dermatology and deep learning with chapters 2 and 3. Chapter 4 presents a review of deep learning applications in dermatology, together with the opportunities and challenges of the field. The anatomical mapping method is introduced in chapter 5, our approach to differential diagnosis in chapter 6 and the disease quantification systems in chapter 7. We continue with the generation of synthetic dermatology images in chapter 8, a presentation of our teledermatology research initiative in chapter 9, a description of the proposed model development framework in chapter 10 and conclude with a final discussion in chapter 11.

Chapter 2

Clinical Dermatology

Dermatology is the branch of medicine studying the skin and its accessory structures, namely hair, nails, sweat glands and sebaceous glands. Experts are called dermatologists. They are medical doctors who followed an additional five years specialized education to understand, diagnose and treat skin disorders. Research in dermatology focuses on understanding the mechanisms underlying skin conditions and develops adapted treatments and surgery procedures.

Skin diseases are highly prevalent. 47.9% of the adult European population suffers from a skin disorder [50]. Society pays a hefty price to treat dermatology patients. In Europe, the costs related to occupational skin diseases alone well exceed five billion euros per year [90]. On the other hand, skin conditions contribute to 1.79% of the global burden of diseases measured in disability-adjusted life years [92, 100] and are rarely lethal. However, they can be physically incapacitating and also have adverse psychological consequences such as depression, anxiety or suicidal ideation [39]. Thus, treatments should always consider both physical and psychological dimensions in patients.

2.1 Morphology of the Skin and its Efflorescences

Efflorescences, also called skin lesions, are changes in the skin resulting from diseases. Before reviewing them, it is useful to first understand the morphology of the skin itself. The skin is composed of three layers, epidermis, dermis, and hypodermis as illustrated in figure 2.1. The epidermis is the outer layer of the skin. It provides a barrier against the outside environment, is waterproof and protects against ultraviolet radiations. The dermis lies beneath the epidermis. It gives the skin its elastic structure and contains the blood vessels, hair follicles and various glands. The innermost layer of the skin is the hypodermis. It is ticker than the other two layers, acts as a shock absorber, contains the fat reserves and helps regulate the body temperature.

The different efflorescences are separated into four main categories: flat lesions, raised solid lesions, fluid-filled lesions and lesions due to broken epidermis. While flat

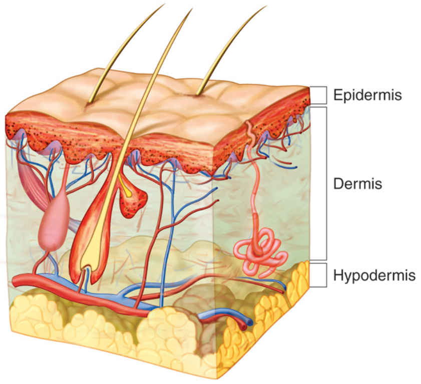


Figure 2.1: Skin morphology. Open Learning Initiative[1], CC BY-NC-SA, no changes.

lesions exclusively affect the epidermis, the others can also affect the dermis, in which case a scar can remain after healing. Dermatologists also differentiate between primary lesions, which are changes appearing directly on healthy skin, and secondary lesions, the subsequent evolution of the latter. The main types of skin lesions are shown in figure 2.2.

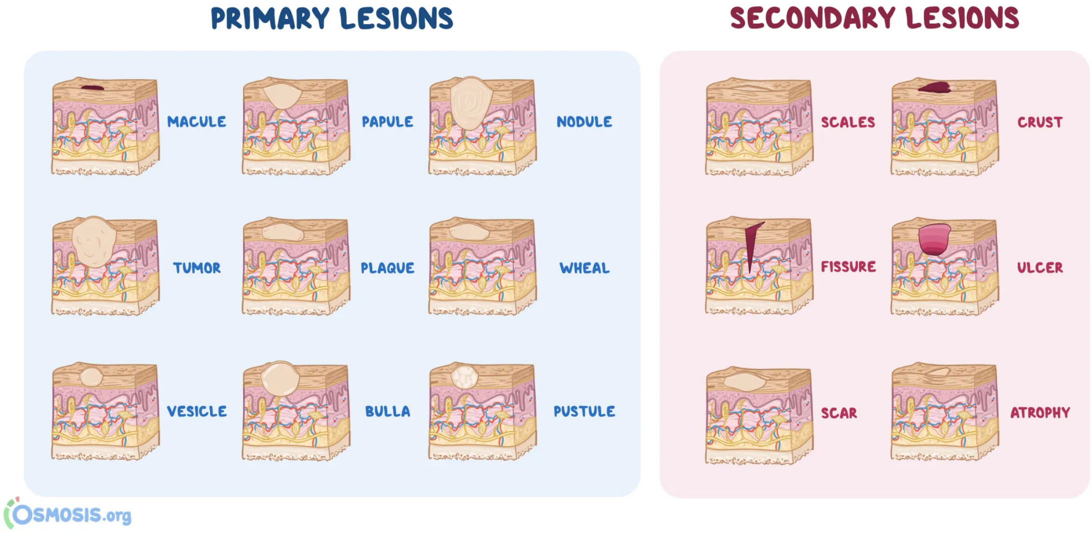


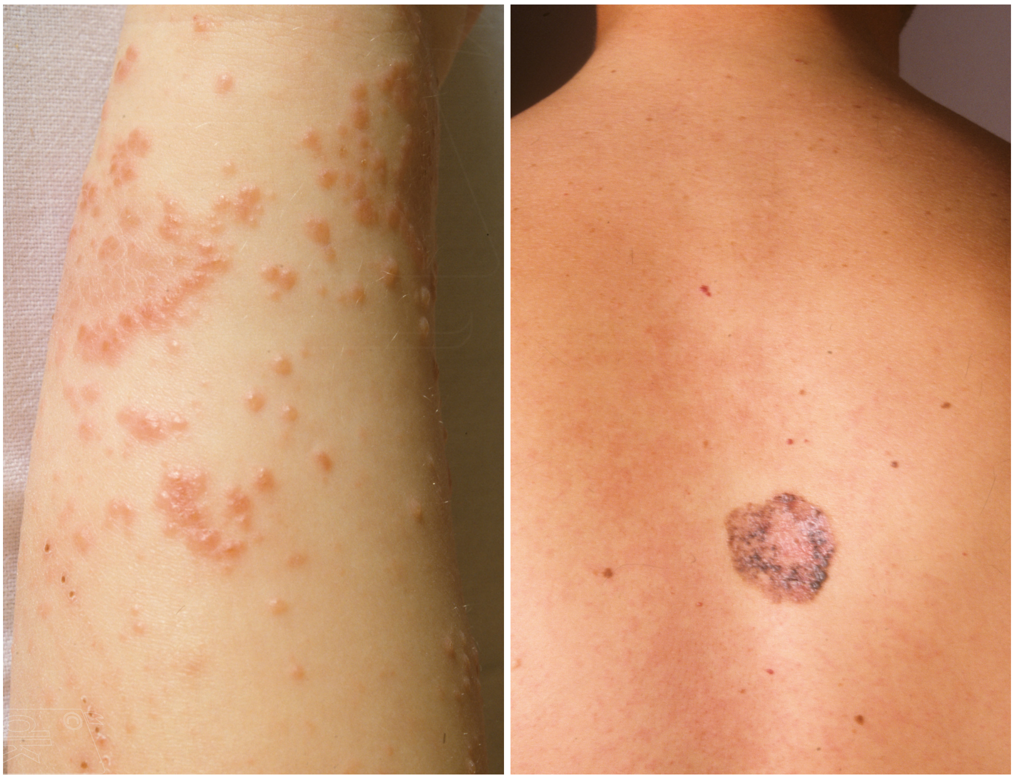
Figure 2.2: Lesions morphology. Osmosis [13], CC BY-NC-ND, no changes.

[1] Image source: `https://www.coursehero.com/study-guides/cuny-csi-ap-1/integumentary-structures-and-functions/` (Accessed: 2nd February 2023)

2.2 Differential Diagnosis of Skin Diseases

There exist over two thousand different skin disorders, the most prevalent in Europe being fungal skin infections, eczema, alopecia, and acne [50]. To diagnose them, the first step is to understand the anamnesis or, in other words, the history of patients. This involves learning how and under what conditions the lesions appeared and how they evolved since then. It is also necessary to know about patients' family and social environment, their past health record and the eventual treatments used.

The second step consists in describing lesions following a standardized procedure to collect all relevant features: location, count, distribution, arrangement, consistency, morphology, texture, color, shape, and appearance of lesions. For example, the lesions in figure 2.3 would be described by table 2.1.



(a) Gianotti-Crosti syndrome

(b) Basal cell carcinoma

Figure 2.3: Examples of skin lesions. DermaCompass[2], CC BY-NC-SA, no changes.

Based on the observed features, dermatologists follow appropriate decision trees to make the differential diagnosis. In unclear or high-risk cases, the diagnosis is verified by taking skin biopsies, which are analyzed in laboratory by a dermatopathologist. To learn and train this process, dermatologists are exposed during their education to sev-

[2] Image source: `https://www.dermacompass.net` (Accessed: 2nd February 2023)

Figure	Location	Number	Distribution	arrangement	consistency
2.3a	lower legs	multiple	localized	-	soft
2.3b	trunk	single	localized	nummular	hard

Figure	morphology	texture	color	shape	appearance
2.3a	papule	smooth	light red	round	symmetrical
2.3b	tumor, crust	crusty	erythema, brown	round	asymmetrical

Table 2.1: Description of skin lesions from figure 2.3.

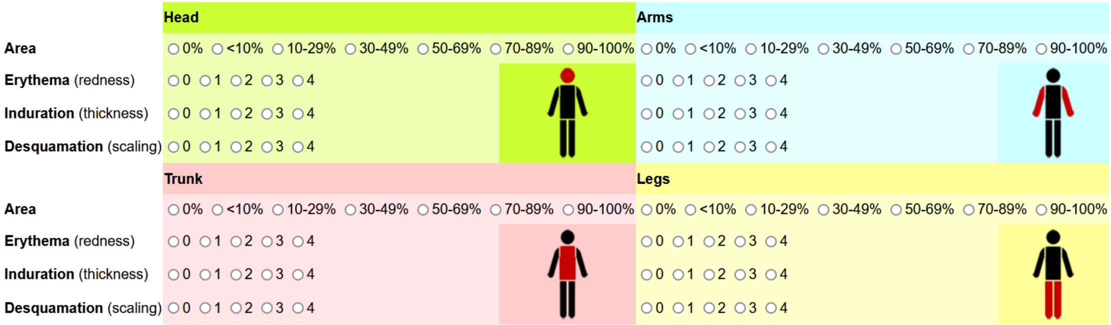
eral thousands of lesion pictures from as many conditions as possible. They are confronted with typical as well as rarer patient cases that could be encountered in practice. Proficiency comes with practice and experience, making the whole process dependent on dermatologists' individual skills.

2.3 Severity Grading of Skin Diseases

After a disease has been diagnosed, dermatologists evaluate its progression and severity. There is no generic severity grading system since conditions have different phenotypes. Disease-specific scoring systems have been proposed instead, enabling dermatologists to make more objective assessments than if they relied on their personal evaluation. Another advantage is that the evolution of patients' conditions can be quantified over time more robustly, allowing to compare the efficacy of different treatments.

Scoring systems usually rely on disease-specific clinical signs and efflorescence quantification, such as lesion counts or surface estimates. For example, the severity of vitiligo is evaluated by estimating the size of lesions with hands surface units, one unit corresponding roughly to one percent of the body, and depigmentation in percentage [71]. Another well-known system is the psoriasis area and severity index (PASI) [56] (cf. figure 2.4). This score considers the four main body regions (head, upper limbs, trunk, lower limbs) and evaluates the skin surface covered by psoriasis from 1 (below 10%) to 6 (above 90%). Each region is graded separately from 0 (none) to 4 (very severe) for erythema, induration, and desquamation (psoriasis clinical signs). The final PASI score is a weighted average of the evaluated ratings.

Although these systems have advantages and benefits in practice, they also present several challenges. More complex scoring systems like PASI necessitate training and experience to be performed efficiently and reliability. They usually cannot be performed by physician extenders, requiring dermatologists to spend part of their consultations on this time-consuming task. Another issue is that these methods remain subjective and suffer from inter-observer and even intra-observer variability [18, 66, 210, 219]. In other words, two dermatologists (or even the same dermatologist after some time) may not necessarily score the same patient equivalently. Finally, these systems remain relatively imprecise due to the scores' discrete nature. For example, two patients, both graded as



	Head	Arms
Area	○0% ○<10% ○10-29% ○30-49% ○50-69% ○70-89% ○90-100%	○0% ○<10% ○10-29% ○30-49% ○50-69% ○70-89% ○90-100%
Erythema (redness)	○0 ○1 ○2 ○3 ○4	○0 ○1 ○2 ○3 ○4
Induration (thickness)	○0 ○1 ○2 ○3 ○4	○0 ○1 ○2 ○3 ○4
Desquamation (scaling)	○0 ○1 ○2 ○3 ○4	○0 ○1 ○2 ○3 ○4
	Trunk	**Legs**
Area	○0% ○<10% ○10-29% ○30-49% ○50-69% ○70-89% ○90-100%	○0% ○<10% ○10-29% ○30-49% ○50-69% ○70-89% ○90-100%
Erythema (redness)	○0 ○1 ○2 ○3 ○4	○0 ○1 ○2 ○3 ○4
Induration (thickness)	○0 ○1 ○2 ○3 ○4	○0 ○1 ○2 ○3 ○4
Desquamation (scaling)	○0 ○1 ○2 ○3 ○4	○0 ○1 ○2 ○3 ○4

Figure 2.4: Psoriasis Area Severity Index Calculator [128], GNU GPL, no changes.

mild for the same disease, may still present notable differences. While this may be sufficient to decide on a treatment recommendation in clinical settings, it is too coarse for precise disease evolution monitoring, drug development experiments or personalized medicine.

2.4 Treatment of Skin Diseases

After diagnosis and severity grading, dermatologists can decide whether treatment is advisable and, in the positive case, make a prescription. They should also consider psychological aspects on top of the disease's physical burden. When needed, psychiatric treatment should also be recommended. There are three main kinds of standard treatments in dermatology: topical, systemic and physical.

Topical treatments are applied to specific locations on the body, usually the lesion's site. They consist of two components: the base (lotions, pastes, ointments) and an active agent. While the latter is chosen according to the disease characteristics (e.g. steroids against inflammation), the base is chosen mainly depending on the hydration of the skin. For example, ointments will be applied on dry skin while lotions will be used on wet lesions. Thus, topical treatments can take various forms from ointments, creams, gels to lotions and pastes.

Systemic treatments target the whole body. They are used when the affected regions are either too large or not easily accessible, e.g. psoriasis and lesions on the scalp. Dermatologists also resort to systemic treatments when topical treatments are ineffective or when a disease has systemic effects. Examples of such treatments are antibiotics, antihistamines and antiviral agents. When prescribing systemic treatments, it is important to consider side effects, as they are more likely to cause undesired complications than topical treatments.

Physical treatments use physical means to treat skin lesions. For example, cryotherapy freezes the skin to induce necrosis of the abnormal lesions, iontophoresis can be used to reduce excessive sweating and phototherapy uses lights with specific wavelengths to treat certain inflammatory skin diseases. Like with any treatment, it is important to determine first the correct diagnosis, if needed by taking a biopsy, before deciding on the appropriate therapy.

Aside from the inherent development and prescription challenges, dermatologists also face the lack of patient adherence to their treatment. In the case of acne therapy, a worldwide observational study reported an adherence rate of only 50% [46]. Adherence varies depending on the skin condition, but it was observed overall that topical treatments' adherence was poorer than systemic treatments [5]. Adherence is an important topic to consider both in the prescription and development of treatments.

2.5 Imaging Modalities in Dermatology

Imaging is widely used in dermatology for teaching and patient documentation [81]. It is an efficient mean to archive and exchange information, both for patients and dermatologists. Imaging makes skin disease monitoring simpler and more precise, as changes can be compared between exact timestamped snapshots of the lesions. This is particularly useful for skin conditions such as skin cancer, where the evolution of a mole can be a strong indicator for malignancy [137]. Precise comparison requires that pictures are properly standardized. Depending on the imaging technique, standardization is either "built-in" or needs to be enforced with protocols [80]. One regulatory concern with imaging is the need for patients' consent, similar to other kinds of medical data. This is particularly important with non-anonymous images, especially when data should be shared or used for research. Another challenge lies in the organization of collected pictures in medical databases and the creation of annotations and metadata necessary for research. While technical platforms are made available by commercial firms, the creation of metadata requires dermatologists' intervention, which is often not possible in practice due to the associated costs.

The most common modality in dermatology is medical photography with standard cameras (e.g. figure 2.3). It is simple to perform, does not require specialized or expensive hardware, and the produced pictures are easily shareable. A survey including 153 board-certified dermatologists from the US reported that 61.8% used this modality every day [129]. The simplicity of medical photography comes with the disadvantage that captured pictures are often poorly standardized, for example in terms of patient posture, captured anatomical region, camera distance, and zoom level. The same survey reported that only 23.7% of dermatologists followed a predefined photography protocol. This modality can also be used by patients themselves with mobile phone cameras, making it particularly suited for teledermatology. However, picture standardization is often much worse than in clinical settings.

A recent extension of medical photography is full-body imaging, which can be performed with commercial products. This modality consists in taking multiple photographs of the patient from different angles using several cameras placed at predefined positions. The acquired pictures are then stitched together either as a panorama or 3D model of the patient body, enabling easier interaction with the data. These products have the advantage to produce relatively standardized images.

Dermatoscopy, also known as dermoscopy, is another modality commonly used in dermatology especially for the diagnosis of skin cancer but also for non-pigmented skin disorders and inflammatory diseases. A dermatoscope or dermoscope is a simple device, which allows to locally magnify lesions (usually tenfold) such as moles with standardized lighting conditions using polarized or unpolarized lights in real time. Studies have shown that dermatologists' differential diagnosis performance for skin cancer improved with this modality [204]. However, it requires specific training to be properly interpreted. Thus, while dermoscopes are accessible and simple to use (they also exist as mobile phone accessories), they are mainly used by clinicians. The produced images are standardized, but may vary depending on the device manufacturer.

Other modalities are less accessible and require specific training, which restricts their use to clinics. Ultraviolet reflectance imaging is useful to assess superficial cutaneous infections and changes in pigmentation [135]. It creates new contrasts and helps with the analysis of skin conditions. Reflectance confocal microscopy can image the skin until superficial dermis with subcellular resolution using near-infrared light. Together with dermoscopy, this modality was shown to improve diagnosis accuracy of skin cancer [55, 116]. Optical coherence tomography can image the skin up to 1.5 mm depth in real time. It is also useful to diagnose skin cancer and can help assess other disorders such as burns or ultraviolet damages.

Finally, an important modality is the imaging of histology slides, which show the microscopic anatomy of biological tissues (e.g. figure 2.5). They are produced using tissue scanners reaching micron-level resolution. These devices produce extremely high-resolution images of several gigabytes each. Their appearance depends on the acquisition procedures and chosen stainings. For several skin conditions, the differential diagnosis clinical gold standard is to take lesion biopsies and create the corresponding histology slides for histopathology analysis.

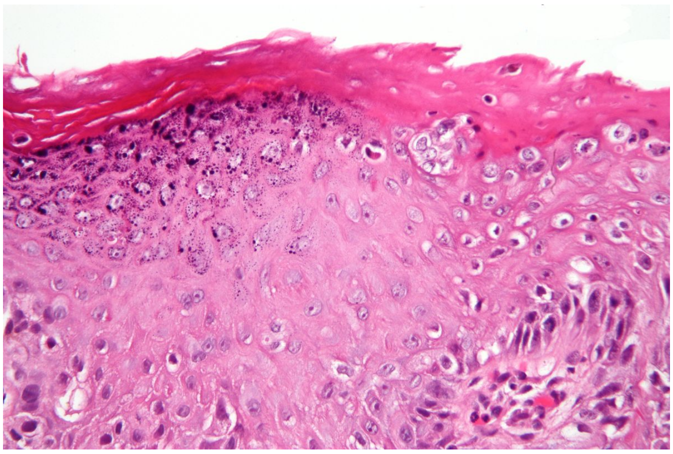


Figure 2.5: Histology slide of a patient with extramammary Paget disease, Michael Bonert [16], CC BY-SA, no changes.

Chapter 3

Deep Learning

Intelligence could be defined as the faculty of reasoning and abstraction. The idea to mechanize intelligence and endow tools or machines with this capacity goes all the way back to antiquity, with mentions already in Greek mythology. Philosophers from many countries and times have attempted to formalize reasoning, giving birth to various logical systems and frameworks. An intuition behind artificial intelligence (AI) is that if reasoning could be reduced to operations over abstract symbols, it could theoretically be performed by machines. In its modern history, AI research aimed at different yet overlapping objectives [169]: the development of agents with human-like behavior and thought process and the development of rational agents able to reason and act logically with respect to an objective. While the latter is based on mathematics and engineering, the former also leverages findings from empirical psychological and biological sciences, resulting in intuitions and heuristics useful for both branches. Today, rational agent approaches such as machine learning (ML) and its sub-field deep learning (DL) are the most successful in terms of practical applications, although they are only capable of restricted tasks and thus belong to weak AI (with respect to general AI on par with human intelligence). In this chapter, we introduce the main concepts of ML and DL with a focus on topics relevant for image-based applications in dermatology.

3.1 Machine Learning Concepts

Machine learning studies algorithms that autonomously learn to perform tasks from data. In this context, an algorithm is considered to "learn" if its performance improves as it processes data. Instead of manually defining specific steps and procedures, these algorithms are designed to automatically extract patterns and statistical relationships from data that are useful to complete the task. The term "learn" is used because the exact operations required may be impractical to define explicitly or even unknown.

For example, dermatologists have methods to screen moles for skin cancer. However, automation with traditional algorithms is infeasible because of the large variety of moles' phenotypes and the difficulty to translate dermatologists' workflow into machine instructions. For this task, ML algorithms are more suitable and were even able to achieve performance on par with human experts [54]. In this example, the algorithm "learned" as long as its malignancy sensitivity increased while it was training over pictures. This learning phase is called the training phase.

3.1.1 Training in Machine Learning

Let us consider a task $t : X \rightarrow Y$ with X the input data domain, Y the corresponding output domain and let us define $\tilde{X} \subseteq X$ as the available training data. Typical ML algorithms consist of two parts, the training and the prediction procedures.

The training procedure optimizes the parameters θ of the algorithm's model m_θ with respect to a performance measure P, which for a given $x \in X$, quantifies how well $m_\theta(x)$ matches $t(x)$:

$$\theta^* = \min_{\theta} \sum_{\forall x \in \tilde{X}} P(m_\theta(x), t(x)) \tag{3.1}$$

The information extracted from the data $\tilde{X}$ is encoded within θ^*, which minimizes the difference as measured by P between m_{θ^*} and t over $\tilde{X}$. The optimization process is specific to the algorithm and should be guided by a performance metric suitable for task t.

Theoretically, if the task is feasible, sufficient training samples are available and the chosen model has enough capacity, it will be able to approximate t. When the model is too complex, the risk is to end up learning training samples, reproducing $\tilde{X}$ perfectly, and to perform poorly on new samples from $X \setminus \tilde{X}$. This phenomenon is called overfitting and can be mitigated by using models with lower capacities or with regularization approaches to ensure that the optimization problem is solved with simple and flexible solutions. The opposite situation is called underfitting. It happens when the model is too simple to learn any useful patterns from the data and performs poorly in general. In these cases, a model with larger capacity is required.

The prediction procedure simply applies the model m_θ to a sample $x \in X$ producing $m_\theta(x)$ as output. While specific algorithms may expect different kinds of data format, they are designed to be agnostic to what the data represents, which makes their application range very wide from medical to financial applications. Thus, the same algorithm can produce models able to solve various tasks, the only difference being the data used, and the parameters learned during training.

3.1.2 Data in Machine Learning

In practice, the success of the training phase depends for a large proportion on the data available to train the model. Usually, only limited data is available, so θ^* may be optimal on $\tilde{X}$ but suboptimal on X. This is usually observed by measuring poor performance when the model is applied on new unseen data, in which case we say that the model does not generalize well. To prevent this, $\tilde{X}$ should be as representative as possible of X with the same underlying data distribution. For example, if a differential diagnosis model is trained on pictures from skin lesions collected in hospitals, we can expect its performance to remain stable for future hospital patients. However, there are no guarantees if we start using this model outside the hospital. In this case, $\tilde{X}$ corresponds to patients visiting hospitals with a disease distribution certainly different from the general population.

The typical ML training workflow starts by splitting the available data $\tilde{X}$ in three sets: the training $\tilde{X}_{train}$, validation $\tilde{X}_{val}$ and testing $\tilde{X}_{test}$ sets. Here again, splitting should be performed in a way ensuring that all three sets remain as representative of X as possible. Then the model is trained on $\tilde{X}_{train}$ while the performance is evaluated on $\tilde{X}_{val}$. Typical ML algorithms have hyperparameters that are optimized by comparing the measured performance on $\tilde{X}_{val}$. Once the training procedure is completed, m_{θ^*} is "frozen" and applied on $\tilde{X}_{test}$ to evaluate the actual performance. $\tilde{X}_{test}$ must contain new data samples different from the ones used to train the model. In clinical settings, it is usually necessary to ensure that a patient's data is only present in one of the splits.

3.1.3 Supervised and Unsupervised Learning

The selection between ML algorithms depends on the available training data and the task to be automated. There are two main families of algorithms used in dermatology applications, which can be defined based on the kind of data available.

Supervised learning concerns situations where both the input data and the task's output data are available. The fundamental tasks in this context are classification where the input should be mapped to a finite set of categories (e.g. benign versus malignant moles) and regression where the input is mapped to a continuous range (e.g. predicting the weight of a patient). Other types of tasks are usually based on either of the two or both, such as learning similarity or ranking. Acquiring both input and output data is often challenging and expensive, especially for tasks in the medical domain, where only experts can create the necessary labels. A subtype of supervised algorithms called active learning focuses on optimizing, which data samples should be labeled to balance this cost while still being able to train models efficiently.

Unsupervised learning refers to algorithms designed for situations where the available data does not contain any output targets. In this context, the algorithms' aim is to find an underlying structure allowing to organize, refine or summarize data. Typical tasks involve clustering, denoising, dimension reduction, anomalies, and outliers identification. In practice, the available data usually comprises a mix of both labeled and unlabeled samples, the latter being often more numerous. This context is called semi-supervised learning and is usually addressed by a combination of algorithms from both categories.

The remaining context, notable for being one of the three basic ML paradigms together with supervised and unsupervised learning, is reinforcement learning. Algorithms in this field aim to train intelligent agents to interact with an environment and find the optimal behavior maximizing a reward scheme. There are many practical applications, especially in robotics.

3.2 Deep Learning Concepts

Deep learning is a sub-field of machine learning studying both supervised and unsupervised algorithms based on the artificial neural network (ANN) architecture. All considerations discussed for ML are also valid in the context of deep learning (DL). While ML algorithms require structured data, typically in the form of handcrafted features, DL algorithms can autonomously learn hierarchies of features from unstructured data (e.g. plain text, images, video, or audio). The underlying drawback being that the latter often require much more data than the former to converge.

Although the inception of ANN algorithms dates back to the 1940s, it was only in 1989 that a practical training method [168] was established based on the backpropagation algorithm [109]. ANN remained rarely used due to their computational costs and the competition of other ML algorithms until an inflection point in 2009, when graphical processing units made training practical, enabling breakthrough results in machine vision and other fields. This led to a revolution of DL with the development of countless ANN architectures customized for different types of applications and data.

3.2.1 Artificial Neural Networks

ANNs are derived from the human brain's organization. Their base building block is the artificial neuron, a mathematical model of biological neurons.

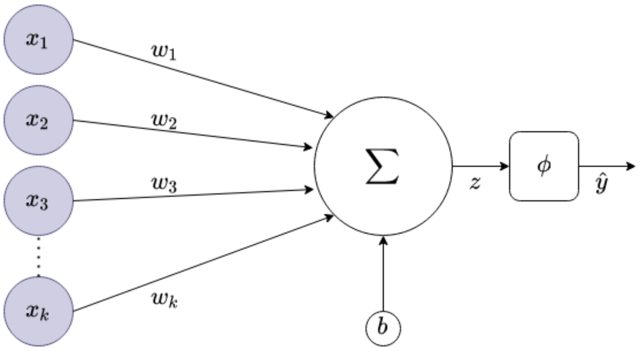


Figure 3.1: Schema of an artificial neuron. The grey circles represent the input layer.

An artificial neuron (figure 3.1), as inspired from the definition of Minsky and Papert [132], takes an input $x \in \mathbb{R}^k$, performs a weighted sum of its dimensions with weights $w_i \in \mathbb{R}$ for $1 \leq i \leq k$, adds a bias term $b \in \mathbb{R}$ and passes the resulting logits $z \in \mathbb{R}$ to a non-linear activation function $\phi : \mathbb{R} \to Y$ mapping them to a range $Y \subseteq \mathbb{R}$:

$$\hat{y} = \phi(z) = \phi(\sum_{i=1}^{k} w_i x_i + b)$$

An ANN is built from several interconnected artificial neurons and can be represented by a directed graph where the nodes are neurons and the edges are connections.

There are two main types of ANNs depending on the nature of these connections: feed-forward and recurrent networks. Feed-forward ANNs comprise all networks designed as directed acyclic graphs, i.e. there are no connections between neurons forming loops. Recurrent ANNs are networks containing loops, which enable them to learn sequential data such as time series or natural language, their internal states serving as memory. Recurrent networks are rarely used in dermatology applications, mainly because sequential data is seldom available. For this reason, we will focus on feed-forward ANNs in this thesis.

Multi-layer Perceptron

In practice, ANNs are built by stacking disconnected neurons in layers. Connections depend on the layer type, the most common being dense layers (also called fully connected layers) where each neuron is connected to every input's element. This layer does not introduce loops, so a network built only from dense layers is a feed-forward network.

A single-layer perceptron is a network built from one dense layer containing a single neuron. It can perform binary classification and find a linearly separable pattern in the training data [132]. For more complex patterns, additional layers with multiple neurons can be stacked one after the other (hence the name "deep" learning) such that a layer's output becomes the following layer's input. The first and last layers of a network are called the input and output layers, while layers in-between are called hidden layers. A network with several dense layers is called a multi-layer perceptron and can theoretically learn any continuous real function according to the universal approximation theorem [82]. Learning is hierarchical, with the successive layers extracting features of increasing complexity. The first layers extract basic features, which are then combined into more complex features that determine the network's output.

In figure 3.2 we present an example multi-layer perceptron with two dense layers of respectively c and c' neurons. Taking $\mathbf{x} \in \mathbb{R}^k$ as the network input, $\mathbf{W}^{(1)} \in \mathbb{R}^{k \times c}$ and $\mathbf{b}^{(1)} \in \mathbb{R}^c$ as the weights and biases of the first hidden layer's neurons, their output can be written as:

$$\mathbf{h}^{(1)} = \phi^{(1)}(\mathbf{z}^{(1)}) = \phi^{(1)}({\mathbf{W}^{(1)}}^\top \mathbf{x} + \mathbf{b}^{(1)})$$

Similarly with $\mathbf{W}^{(2)} \in \mathbb{R}^{c \times c'}$ and $\mathbf{b}^{(2)} \in \mathbb{R}^{c'}$ the weights and biases of the second hidden layer, the output of the network is:

$$\hat{\mathbf{y}} = \phi^{(2)}({\mathbf{W}^{(2)}}^\top \mathbf{h}^{(1)} + \mathbf{b}^{(2)}) \tag{3.2}$$

Where $\phi^{(i)}$ is the activation function of layer i.

Activation Functions

Without activation functions, the operations of an ANN could be reduced to a single linear transformation since artificial neurons would only compute the weighted sum of their input. Activation functions are thus an essential component, enabling ANNs to

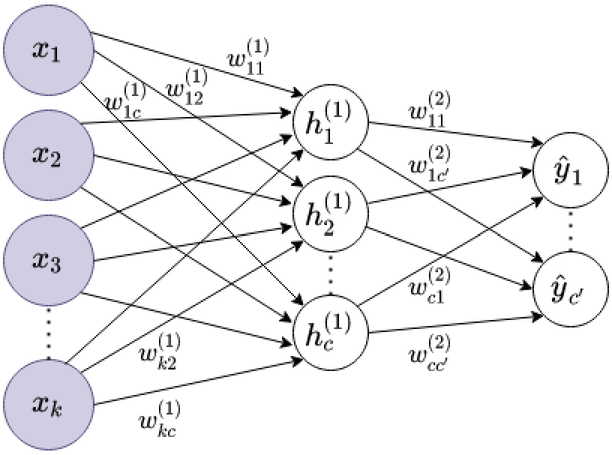


Figure 3.2: Schema of a two layers multi-layer perceptron. The white circles represent full neurons (including bias and activation function) as described in figure 3.1. The subscript indexes are composed of two numbers identifying respectively the source and destination, while the exponent indexes in parentheses are the layer identifiers.

learn non-linear patterns in data (when the activation functions are non-linear themselves). The main requirements for activation functions are that they should be efficient computationally, continuous, differentiable and produce informative gradients for the learning process. Several functions have been proposed to meet these objectives. We present in the following some of the most frequently used in practice.

Sigmoid The sigmoid function maps logits between 0 and 1, which can be interpreted as a probability distribution. For this reason, it is often used at the output layer in binary or multi-label classification tasks. An issue faced with the sigmoid function is saturation for large input logits, i.e. the output values are close to the asymptotic bounds of the function leading to the vanishing gradients problem.

$$\sigma(z) = \frac{1}{1 + \exp\{-z\}}$$

$$\frac{\partial \sigma(z)}{\partial z} = \sigma(z)(1 - \sigma(z))$$

ReLU The rectified linear unit [141] clips negative logits to zero. It is efficient computationally and mitigates the vanishing gradients problem as it does not saturate. However, it has the drawback to produce zero gradients for negative logits (called the dying units problem).

$$\begin{aligned}
\mathrm{ReLU}(z) &= \mathrm{max}(0, z) \\
\frac{\partial \mathrm{ReLU}(z)}{\partial z} &= \begin{cases} 0, & \text{if } z < 0 \\ 1, & \text{otherwise} \end{cases}
\end{aligned}$$

ELU The exponential linear unit [37] activation function is an extension of ReLU, which mitigates the dying units problem by ensuring that gradients remain non-zero for negative logits. Given a positive real hyperparameter α, it is defined as follows:

$$\begin{aligned}
\mathrm{ELU}_\alpha(z) &= \begin{cases} z, & \text{if } z > 0 \\ \alpha(\exp\{z\} - 1), & \text{otherwise} \end{cases} \\
\frac{\partial \mathrm{ELU}_\alpha(z)}{\partial z} &= \begin{cases} 1, & \text{if } z > 0 \\ \alpha \exp\{z\}, & \text{otherwise} \end{cases}
\end{aligned}$$

Swish The swish activation function [159] is based on the sigmoid function and a learnable parameter β. It does not cause the dying units problem and has the property to be non-monotonic.

$$\begin{aligned}
\mathrm{Swish}_\beta(z) &= z\sigma(\beta z) \\
\frac{\partial \mathrm{Swish}_\beta(z)}{\partial z} &= \beta \mathrm{Swish}_\beta(z) + \sigma(\beta z)(1 - \beta \mathrm{Swish}_\beta(z))
\end{aligned}$$

Softmax The softmax function is mostly used on the output layer for multi-class classification tasks. Whereas other activation functions perform element-wise operations, the softmax function produces a probability distribution from all layer's logits taken together. Given a layer's output logits $\mathbf{z} \in \mathbb{R}^N$, the softmax function is computed as follows:

$$\begin{aligned}
S(\mathbf{z}, i) &= \frac{\exp(z_i)}{\sum_{j=1}^{N} \exp\{z_j\}} \\
\frac{\partial S(\mathbf{z}, i)}{\partial z_j} &= \begin{cases} S(\mathbf{z}, i)(1 - S(\mathbf{z}, j)), & \text{if } i = j \\ -S(\mathbf{z}, i)S(\mathbf{z}, j), & \text{otherwise} \end{cases}
\end{aligned}$$

3.2.2 Training Artificial Neural Networks

ANNs are trained following the ML training paradigm (cf. section 3.1.1), based on two key elements: a loss function as the performance measure to be optimized and a procedure to update the network's parameters. In deep learning, the most common update procedures are variations of the gradient descent algorithm, which is based on the computation of gradients determined using the backpropagation algorithm [109, 168]. The training procedure consists in optimizing the network's weights such that the loss function is minimized given the training samples.

Loss Functions

The aim of loss functions (also called objective functions) is to quantify the difference between expected ground truth targets $\mathbf{y}$ and the observed model predictions $\hat{\mathbf{y}}$. The design of a loss function depends on the task to be solved. In the following, we present a selection of frequently used loss functions for classification and segmentation tasks in a context with N training samples and C classes. With $\hat{y}_i^{(c)}$, we denote the predicted probability that data sample i belongs to class c. We use $y_i^{(c)}$ to represent the same for the ground truth.

Cross Entropy Loss This loss function compares the difference between predicted and expected probability distributions of the task's targets.

$$\mathcal{L}_{CE}(\mathbf{y}, \hat{\mathbf{y}}) = -\frac{1}{N}\sum_{i=1}^{N}\sum_{c=1}^{C} y_i^{(c)} \log\left(\hat{y}_i^{(c)}\right)$$

Focal Loss The focal loss [107] extends the cross entropy loss to mitigate class imbalance by focusing on hard training samples based on a hyperparameter $\gamma \geq 0$.

$$\mathcal{L}_{F}(\mathbf{y}, \hat{\mathbf{y}}) = -\frac{1}{N}\sum_{i=1}^{N}\sum_{c=1}^{C} y_i^{(c)}(1 - \hat{y}_i^{(c)})^{\gamma} \log\left(\hat{y}_i^{(c)}\right)$$

Dice loss The dice loss [130] is used mainly in segmentation tasks to measure the overlap between the predictions and ground truth masks.

$$\mathcal{L}_{D}(\mathbf{y}, \hat{\mathbf{y}}) = 1 - 2\frac{\sum_{i=1}^{N}\sum_{c=1}^{C} y_i^{(c)}\hat{y}_i^{(c)}}{\sum_{i=1}^{N}\sum_{c=1}^{C} y_i^{(c)} + \hat{y}_i^{(c)}}$$

Gradient Descent Algorithm

Gradient descent is an optimization algorithm that minimizes a function by iteratively updating its parameters in the opposite direction of its gradients. In the case of ANNs, the algorithm minimizes a loss function with respect to the network's parameters (the weights and biases) over the training samples. In its simplest form, this update at iteration step $t + 1$ is performed by subtracting the loss gradients scaled by a learning rate hyperparameter $\eta > 0$ from the network parameters $\boldsymbol{\theta}_t$ at step t:

$$\boldsymbol{\theta}_{t+1} = \boldsymbol{\theta}_t - \eta\nabla\mathcal{L}(\boldsymbol{\theta}_t)$$

When a function is convex, the algorithm is guaranteed to converge to its global minimum. However, the loss function of an ANN is non-convex due to the non-linearities introduced by the activation functions and convergence may occur at a local minimum. To improve convergence rate and reduce the tendency to remain at local minima, more complex update rules were proposed, mainly based on statistics of gradients from previous iteration steps.

Momentum This approach [186] consists in updating the network's parameters based on both the loss gradients at step t and the exponential moving average of the first moment of previous iterations' gradients. The percentage of retained past gradients is controlled with a hyperparameter β. The intuition is to keep track of the general downward direction of the loss surface and not get stuck in flat regions or local minima.

$$\begin{aligned}
\mathbf{m}_{t+1} &= \beta \mathbf{m}_t + \eta \nabla \mathcal{L}(\theta_t) \\
\theta_{t+1} &= \theta_t - \mathbf{m}_{t+1}
\end{aligned}$$

Adam The adaptive moment estimation optimizer [98] leverages momentum and adjusts the learning rate for each network's parameters based on the combined exponential moving average of the first and second past gradient moments. It uses two hyperparameters β_1 and β_2 to control the decay of respectively the first and second gradient moments.

$$\begin{aligned}
\mathbf{m}_{t+1} &= \beta_1 \mathbf{m}_t + (1-\beta_1)\nabla \mathcal{L}(\theta_t) \\
\mathbf{v}_{t+1} &= \beta_2 \mathbf{v}_t + (1-\beta_2)(\nabla \mathcal{L}(\theta_t))^2 \\
\theta_{t+1} &= \theta_t - \eta \frac{\mathbf{m}_{t+1}/(1-\beta_1^{t+1})}{\sqrt{\mathbf{v}_{t+1}/(1-\beta_2^{t+1})}}
\end{aligned}$$

Backpropagation Algorithm

The key of gradient descent is to efficiently compute the loss function's gradients with respect to every trainable parameters of the ANN. This is achieved with the backpropagation algorithm [109], which operates in two main phases. First, the network predictions are computed in the forward propagation phase (with operations similar to equation 3.2) and used to compute the loss function. Second, the loss gradients are computed. Since the forward propagation can be reduced to a composition of operations, and provided that all operations are differentiable, the chain rule from calculus can be applied to compute the loss partial derivative with respect to each of the network's parameters. Taking the network in figure 3.2 as an example, the loss $\mathcal{L}$ partial derivative with respect to the weight $w_{11}^{(1)}$ is calculated as follows:

$$\frac{\partial \mathcal{L}}{\partial w_{11}^{(1)}} = \left(\sum_{j=1}^{c'} \frac{\partial \mathcal{L}}{\partial \phi^{(2)}} \times \frac{\partial \phi^{(2)}}{\partial z_j^{(2)}} \times \frac{\partial z_j^{(2)}}{\partial \phi^{(1)}} \right) \times \frac{\partial \phi^{(1)}}{\partial z_1^{(1)}} \times \frac{\partial z_1^{(1)}}{\partial w_{11}^{(1)}}$$

Similarly, the loss $\mathcal{L}$ partial derivative with respect to the bias $b_1^{(1)}$ is obtained by:

$$\frac{\partial \mathcal{L}}{\partial b_1^{(1)}} = \left(\sum_{j=1}^{c'} \frac{\partial \mathcal{L}}{\partial \phi^{(2)}} \times \frac{\partial \phi^{(2)}}{\partial z_j^{(2)}} \times \frac{\partial z_j^{(2)}}{\partial \phi^{(1)}} \right) \times \frac{\partial \phi^{(1)}}{\partial z_1^{(1)}} \times \frac{\partial z_1^{(1)}}{\partial b_1^{(1)}}$$

Performance Evaluation Metrics

The evaluation of a network's performance is achieved by comparing its validation or test set predictions with the expected ground truth labels. Several metrics have been devised to perform this comparison objectively. Their choice depends on the target application. In the following, we present a selection of metrics frequently used in the context of image classification and segmentation tasks. Given a class c, we use the notation TP, TN, FP, FN for, respectively, the true positive, true negative, false positive and false negative counts. In this context, true and false correspond to whether a model prediction matches the ground truth, and positive or negative whether a sample belongs to class c.

Accuracy The accuracy measures the proportion of correct predictions. The drawback of this metric is that it will be dominated by the most prevalent classes, which becomes problematic in the context of severe class imbalance. In this case, other metrics should be selected.

$$ACC_c = \frac{TP_c + TN_c}{TP_c + TN_c + FP_c + FN_c}$$

Precision Precision measures the proportion of true positives among all positive predictions.

$$P_c = \frac{TP_c}{TP_c + FP_c}$$

Sensitivity Sensitivity (also called recall or true positive rate) measures the proportion of correct positive predictions among all samples actually belonging to class c.

$$Sen_c = \frac{TP_c}{TP_c + FN_c}$$

Specificity Specificity (also called true negative rate) measures the proportion of correct negative predictions among all samples that do not belong to class c.

$$Spe_c = \frac{TN_c}{TN_c + FP_c}$$

Balanced Accuracy The balanced accuracy mitigates the drawback of plain accuracy and can be used in imbalanced context.

$$BACC_c = \frac{Sen_c + Spe_c}{2}$$

F1 Score The F1 score is the harmonic mean of the precision and sensitivity.

$$F1_c = 2\frac{P_c * Sen_c}{P_c + Sen_c} = \frac{2TP_c}{2TP_c + FP_c + FN_c}$$

Intersection over Union Intersection over union (IoU or Jaccard index) is used in segmentation tasks to evaluate the proportion of overlapping pixels between predictions and ground truth.

$$\mathrm{IoU}_c = \frac{TP_c}{TP_c + FP_c + FN_c}$$

Dice Score The dice score is mainly used in segmentation tasks. It is equivalent to the F1 score and measures the overlap of predicted pixels with the ground truth. The main difference with the intersection over union arises when averaging predictions over multiple samples. In this case, the IoU will penalize poor segmentation of individual samples more than the dice score.

$$Dice_c = F1_c = \frac{2TP_c}{2TP_c + FP_c + FN_c}$$

3.2.3 Convolutional Neural Networks

An issue of dense layers is their exploding computation and memory footprints when applied to high dimensional inputs such as images or videos. For example, to process a 4K image, a dense layer with ten neurons would need to store almost 250 million weights and perform the corresponding number of operations. While images are often resized to much smaller dimensions in practice, learning relevant features from such input data with dense layers remains inefficient. The solution comes from the observation that object detection in images can be broken down into the detection of simpler elements. Objects can be decomposed into parts, which can be divided into shapes, all the way down to simple corners and edges. Since such basic elements do not span over the whole image but occur in delimited regions, it is not necessary that the model considers the full image altogether. Instead, it only needs to process neighboring pixels together and focus on local regions to identify features relevant for the hierarchy of constituent elements composing the objects of interest. It should thus be possible to use only sparse connections, which require less memory and computations. This concept is called local connectivity [65] and can be achieved with the convolution operation. An ANN based on this operation is called a convolutional neural network (CNN). This architecture was first introduced to perform handwritten digits recognition [101].

Convolution Layers

The convolution operation (figure 3.3) takes a kernel $K \in \mathbb{R}^{M \times N}$ with a bias $b \in \mathbb{R}$ and applies them on the input image I to produce a feature map F. For valid i and j values within the image dimensions, this corresponds to the following operation:

$$F(i,j) = (K * I)(i,j) = \sum_{m=1}^{M} \sum_{n=1}^{N} I(i+m, j+n) K(m,n) + b$$

The steps in the image dimensions (successive i, j coordinates) are determined by a hyperparameter called the stride. Furthermore, the image can be padded with zero elements (the number is controlled with a hyperparameter) to accommodate the kernel dimensions, enabling to process the image boundaries.

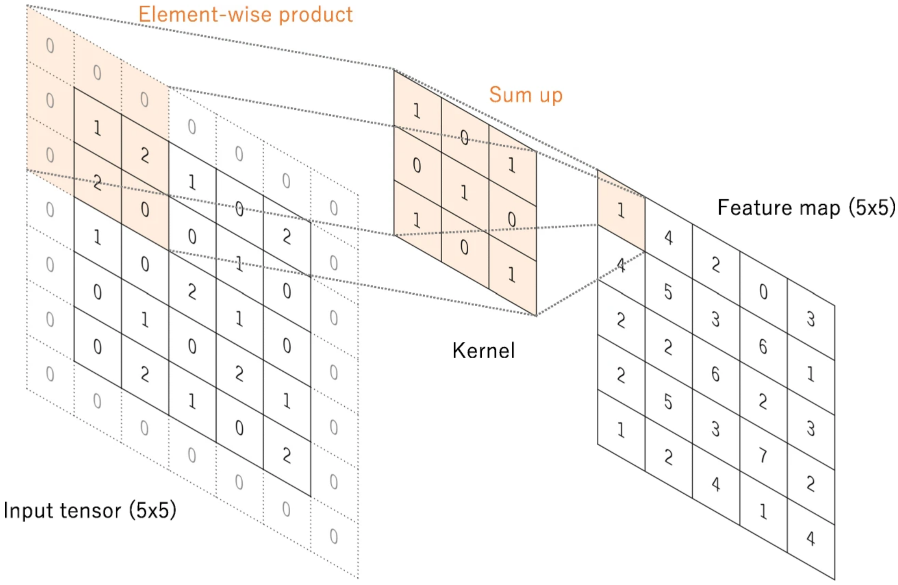


Figure 3.3: Convolution operation with padding 1, 3 × 3 kernel size and stride 1. Yamashita et al. [215] CC BY 4.0, Springer, no changes.

Convolutional layers are layers performing several convolution operations. Since they essentially consist of a list of kernels and biases, the computation and memory requirements are much smaller than dense layers. Kernels are much smaller than the input image. Their size determines the connection of a neuron with its input, where for the first layer, this corresponds to a region in the original image. A larger kernel will connect neurons to a larger region. For subsequent layers, these are regions in the output features of previous layers. Projected back to the input image, these connected regions are called receptive fields. The deeper a neuron is located in a network, the larger grows its receptive field, eventually up to the full image size.

An important property of the convolution operation and convolutional layers is that they are equivariant to translation. In other words, a translated input will result in the same features as if the translation operation was applied on the features extracted from the original untranslated input. In the context of images, moving an object will similarly move the associated features in the feature map.

Pooling Layers

Once features are extracted by convolutional layers, they go through a non-linear activation function and can be aggregated using pooling layers (figure 3.4). The pooling operation consists in summarizing a fixed neighborhood of features, for example by keeping the maximum value. The neighborhood size and stride (steps within the feature map) are predefined hyperparameters. This operation is useful to reduce the features' dimension, thereby decreasing the memory and computation requirements, and as a regularization mean to constrain the network's solution space and prevent overfitting. The side effect of the pooling operation is to make the features approximately invariant (i.e. independent) to small translations on top of the convolution translation equivariance, which becomes problematic if object location is relevant for the task. from another perspective, while the features gain semantic information as the network gets deeper, the pooling operation causes a loss of object location and shape information.

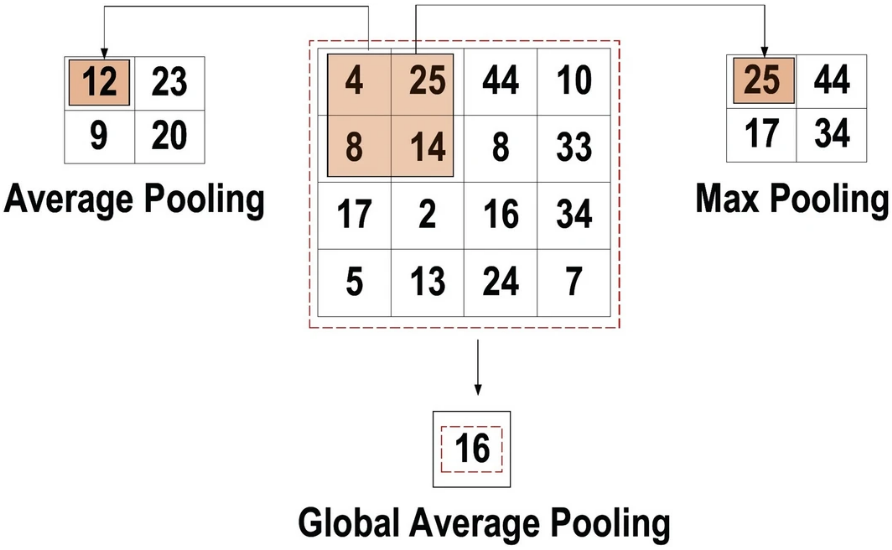


Figure 3.4: Pooling operations. Alzubaidi et al. [8] CC BY 4.0, Springer, no changes.

Transposed Convolutions

While convolutions down-sample the input's dimensions (or keep them unchanged), transposed convolutions (also called deconvolutions) up-sample them, as shown in figure 3.5. They are often used in segmentation tasks to expand the features back to the original size of the input image. Given an input feature map F and a kernel K, the transposed convolution's output will have the same shape (but not values) as the original convolution input I satisfying $F = K * I$.

Atrous Convolutions

Similarly to transposed convolutions, atrous convolutions (also called dilated convolutions) up-sample their input (cf. figure 3.6). This is performed by inserting $d - 1$ zeros in-between kernel elements, where d is a hyperparameter called the dilation rate.

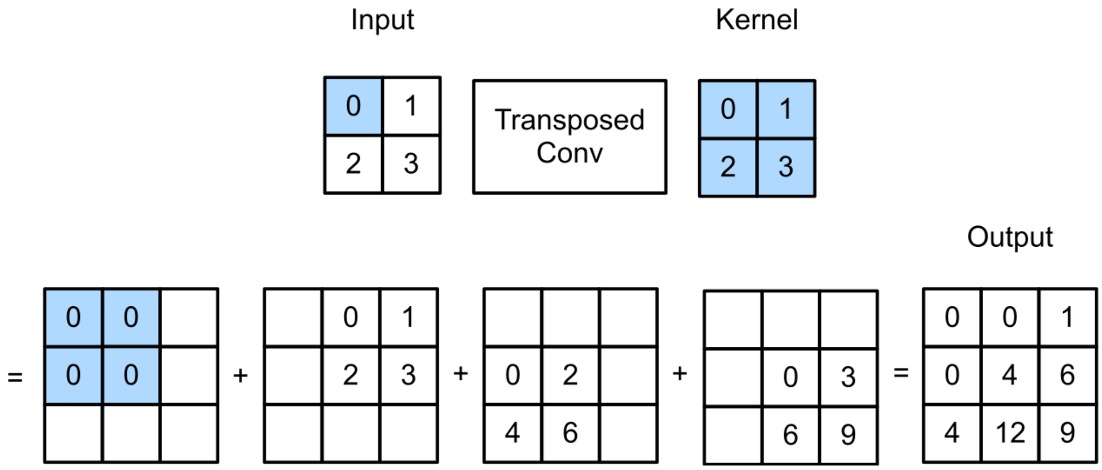


Figure 3.5: Transposed convolution of a 2 × 2 feature map with 2 × 2 kernel with stride 1 and no padding. Zhang et al. [222] CC BY SA 4.0, arXiv, no changes.

This method increases the receptive field of the kernel without additional memory and computational costs.

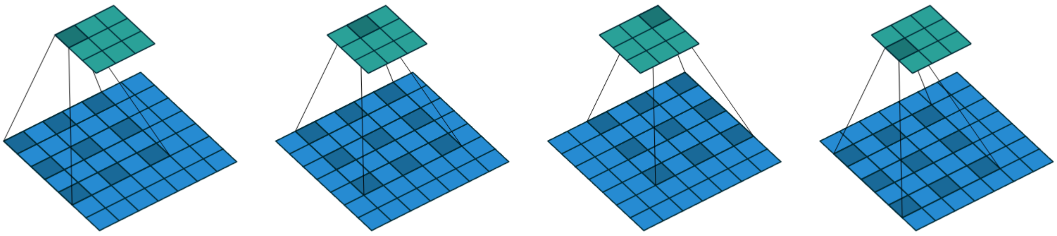


Figure 3.6: Atrous convolutions of a 7 × 7 input with 3 × 3 kernel, dilation factor of 2, stride 1 and no padding. Dumoulin et al. [48] MIT License, arXiv, no changes.

Classification CNNs

Image classification consists in processing an image to predict a category among multiple possibilities. A variation of this task called multi-label classification involves predicting multiple categories per input image. Many CNN architectures have been devised [8, 94] and benchmarked using public datasets such as ImageNet [44]. In the following, we discuss a selection of landmark architectures that introduced important innovations.

VGG The virtual geometry group's network [182] is one of the early very deep ANNs based only on convolutional and pooling operations. The final classification predictions are obtained by passing the extracted feature maps through fully connected layers. VGG popularized the use of small 3×3 kernel sizes and 1×1 convolutions to reduce complexity and computation requirements, although this remains a drawback of this architecture due to its large number of parameters (over 140 millions).

Inception A series of architectures with iterative improvements, starting with Inception v1 (GoogleLeNet) [188], which introduced the inception block (figure 3.7). This block is composed of parallel convolutions with different filter sizes to extract features at different scales. The network was very deep at the time and used intermediate auxiliary classifiers and loss functions to prevent the gradient vanishing problem. In comparison with VGG, this network was optimized to reduce computations with 1×1 convolutions, bottleneck and global average pooling layers, resulting in 5 million parameters.

Inception v2 and v3 [189] introduced several optimizations to decrease further computation requirements, mainly replacing 5×5 convolutions with two 3×3 convolutions and factorizing $n \times n$ convolutions into successive $1 \times n$ and $n \times 1$ convolutions. To reduce representational bottleneck, the authors widened the inception blocks with parallel factorized convolutions. In Inception v3, they improved performance further by introducing factorized 7×7 convolutions, batch normalization and label smoothing.

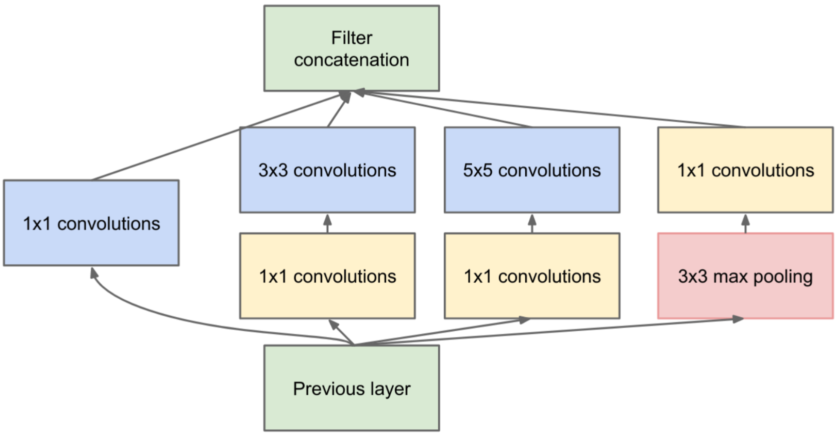


Figure 3.7: Inception block. Szegedy et al. [188] ©2015 IEEE, no changes.

Inception v4 and Inception-ResNet [187] are currently the latest iteration of the architecture. The authors mainly applied various simplifications to the network and changed the stem (initial convolution layers following the input layer) preceding the inception blocks. In Inception-ResNet, they updated the inception blocks to include scaled residual connections inspired from the ResNet architecture [77].

ResNet To alleviate the gradient vanishing problem, the authors [77] introduced residual blocks (figure 3.8), enabling them to train very deep ANNs. These blocks leverage skip connections that allow gradients to directly flow through the network. An extension of ResNet is the ResNeXt architecture [214], which uses parallel residual blocks (figure 3.8) similarly to how the inception network performs parallel convolutions.

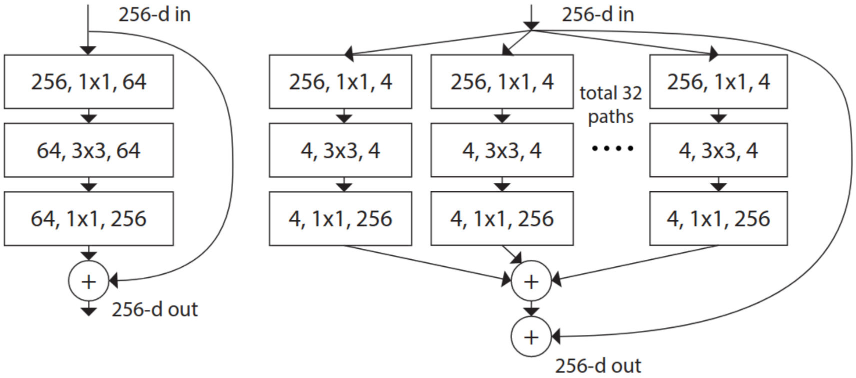


Figure 3.8: Residual block (left) and 32 parallel residual blocks (right). Xie et al. [214] ©2017 IEEE, no changes.

SENet This architecture [86] introduced the squeeze-and-excitation block (figure 3.9), which scales each feature channel with learnable factors and enables the network to weight channels adaptively instead of equivalently. This relatively simple innovation improved performance of ResNet and Inception networks with very little computational overhead.

EfficientNet The development of new ANN architectures typically involves tuning a network's width, depth, and resolution. This process is usually performed according to researchers' empiric intuition rather than systematically. Tan et al. [190] introduced a compound coefficient to methodically scale (figure 3.10) these dimensions in a simple network called EfficientNet-B0, which leverages residual and squeeze-and-excitation blocks. With this approach the authors proposed seven different networks EfficientNet-B1 to EfficientNet-B7 able to process input images of increasing size with improving performance and computation requirements.

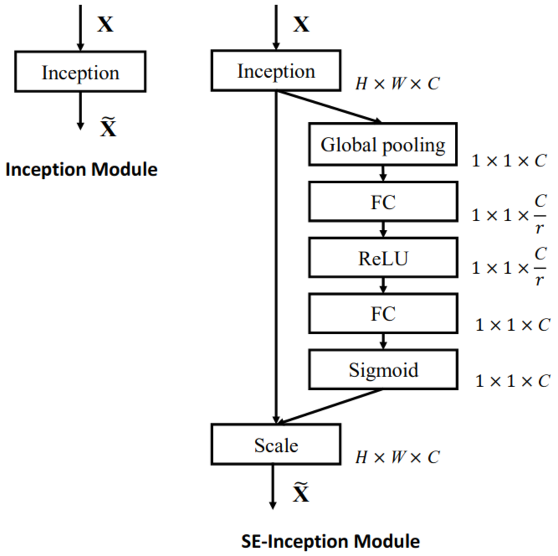


Figure 3.9: Inception block (left) and squeeze-and-excitation inception block (right). Hu et al. [86] ©2018 IEEE, no changes.

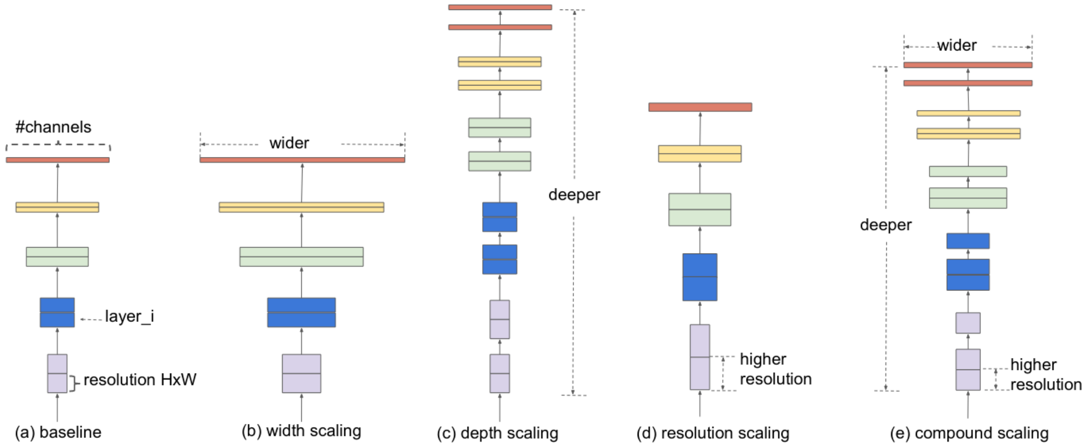


Figure 3.10: EfficientNet's approach to model scaling. Tan et al. [190] ©2019 JMLR, no changes.

Segmentation CNNs

While overall image categories are predicted in classification tasks, semantic segmentation performs pixel-level classification of semantic categories. The result of this operation is called a segmentation mask. Common variations of this task include instance segmentation where distinct objects are segmented separately and panoptic segmentation, a combination of semantic and instance segmentation. Several approaches have been proposed to solve these tasks [8, 131]. They are mostly based on two steps, starting with the extraction of image features, and followed by the expansion of these features to produce the segmentation mask. In the following, we discuss a selection of important segmentation architectures.

FCN Long et al. [115] proposed a fully convolutional architecture applicable to any classification CNN, where the final fully connected layers are replaced with 1 × 1 convolutions. The coarse semantic features extracted by the network are fused with better localized appearance features from a selection of the middle layers (connected via skip connections, see figure 3.11) and then up-sampled with transposed convolutions. The segmentation mask is obtained by applying a final 1 × 1 convolution to adjust the up-sampled features to the number of semantic classes. One downside of this approach is its high computational requirements, which preclude real-time inference.

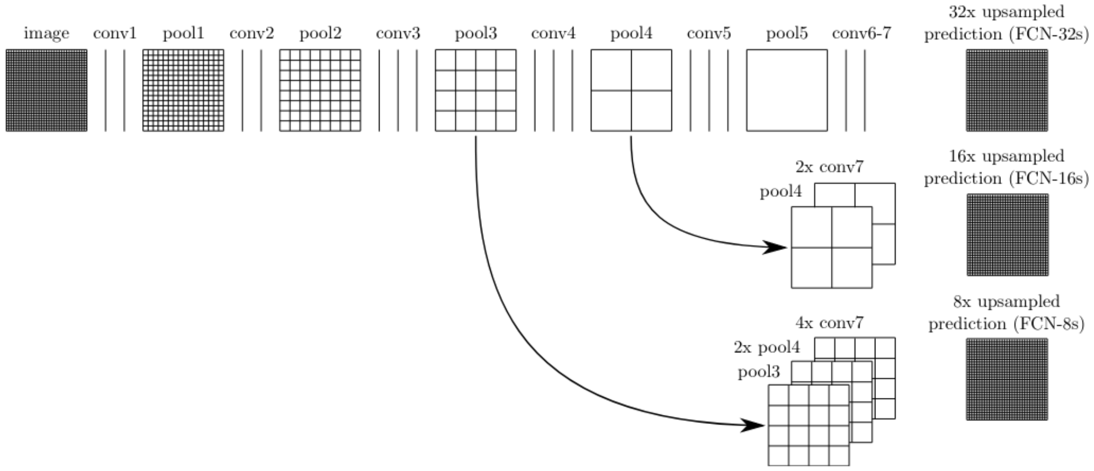


image
conv1
pool1
conv2
pool2
conv3
pool3
conv4
pool4
conv5
pool5
conv6-7
32x upsampled prediction (FCN-32s)
2x conv7
pool4
16x upsampled prediction (FCN-16s)
4x conv7
2x pool4
pool3
8x upsampled prediction (FCN-8s)

Figure 3.11: FCN feature fusion approaches: FCN-32s (first row from top) does not use earlier layers' features, FCN-16s (second row) uses the fourth pooling layer's features, FCN-8s (third row) uses features from both pooling layers 3 and 4. Long et al. [115] ©2015 IEEE, no changes.

U-Net This encoder-decoder architecture was introduced by Ronneberger et al. [164] to segment biomedical images based on small training datasets. First, the encoder iteratively down-samples the input image with convolution and pooling operations to extract compressed context features. Then, the decoder performs symmetric expansion steps to enlarge the features back to the input dimensions, resulting in the final segmentation mask. In the expansion steps, the features are up-sampled, convolved, then concatenated with corresponding down-sampled features from the encoder branch (connected via skip connections) and finally convolved again. The skip connections enable the network to retain appearance information while expanding the features. This architecture can be modified to leverage any classification backbone (CNN stripped down from its final fully connected layers) as encoder provided the decoder is adapted to match the down-sampling steps with corresponding up-sampling steps (e.g. U-Net with ResNet backbone, figure in section 7.1).

Mask R-CNN Based on Faster R-CNN [62, 160], an architecture devised to solve the task of object detection in images, Mask R-CNN [76] performs instance segmentation of the detected objects. The network starts with a CNN backbone, which extracts image features. Then it operates in two phases. First, the region proposal network predicts regions containing objects candidates (regions of interest, RoI). Second, the RoI features are extracted and used to predict the object categories together with their bounding box coordinates. In parallel, the object instance segmentation mask is created by expanding the RoI features with transposed convolutions (see figure 3.12).

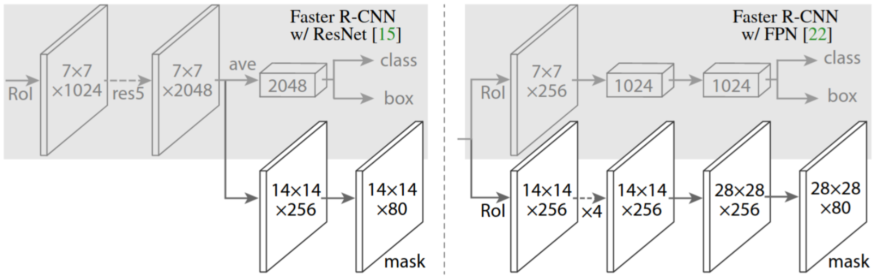


Figure 3.12: The Mask R-CNN's head architectures (left uses a ResNet backbone and right a Feature Pyramid Network backbone [106]). The heads predict the objects' categories, bounding boxes and instance segmentation masks. He et al. [76] ©2017 IEEE, no changes.

DeepLab The first iteration [30] of this architecture used a CNN backbone where the last layer's features were concatenated with features from early layers. The resulting features were then expanded with atrous convolutions and bi-linear interpolation. Furthermore, the authors used fully connected conditional random fields (CRF) to iteratively refine the resulting masks and recover detailed local structures in a post-processing step. This probabilistic graphical model aims to maximize agreement between similar pixels, while leveraging semantic class context.

DeepLabv2 [31] used atrous spatial pyramid pooling (ASSP) as a mean to better segment objects at different scales. This method consists in performing multiple atrous convolutions with different dilation rates and finally pool the results together.

DeepLabv3 [32] up-samples the backbone's image features with cascaded blocks, each of them based on two convolutions and modified ASSP. ASSP was modified by applying 1 × 1 convolution before three atrous convolutions with different dilation rates, appending global average pooled features of the preceding block's features and applying batch normalization. These changes enabled the authors to discard the CRF post-processing step without loss of performance.

DeepLabv3+ [33] is based on an encoder-decoder architecture (cf. figure 3.13). The encoder is similar to DeepLabv3 with a single up-sampling block. To decrease computation costs, convolutions are factored with a depth-wise convolution and a 1 × 1 convolution. Then, the decoder consists in concatenating the 1 × 1 convolved backbone's features with the up-sampled features of the last encoder's block followed by convolutions and another up-sampling operation.

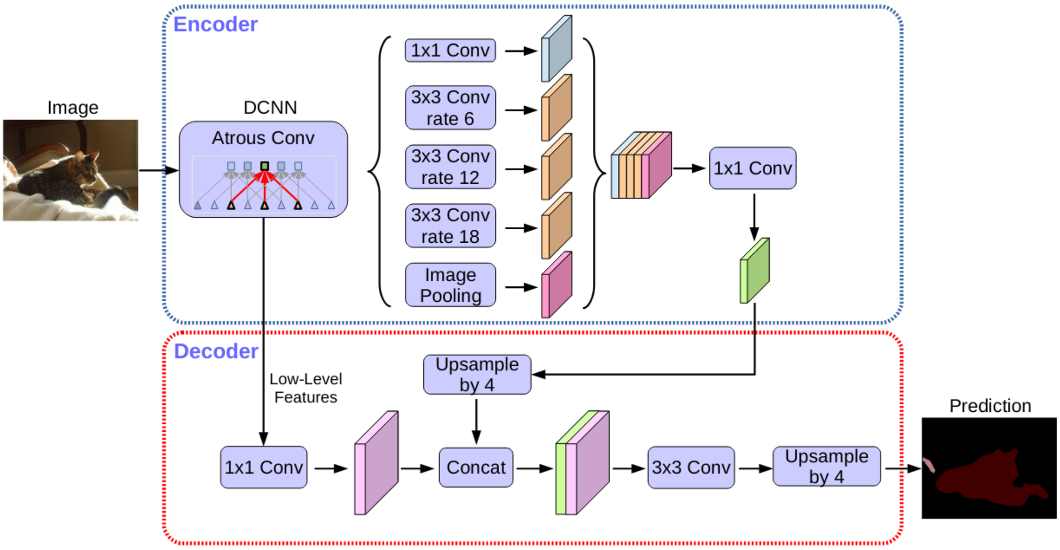


Figure 3.13: DeepLabv3+ architecture. Chen et al. [33] ©2018 IEEE, no changes.

3.2.4 Embeddings

An embedding is a mapping of a discrete input domain X to a continuous vector space E_X called the embedding space. It can be understood as a continuous representation of X encoding relationships between samples that are meaningful in the learning context. Embedding samples enables to evaluate their similarity and to perform operations such as addition, subtractions, or interpolation.

Training ANNs consists in learning such a representation, the features extracted from the data samples being their embedding vectors. Dimension reduction methods can then be applied to visualize embedded data samples, which enables to identify clusters or outliers, and assess how well the ANNs separates the data (e.g. figure 3.14). In most cases, E_X has lower dimension than X especially when considering unstructured data such as text, images, or videos. It can also have higher dimension, for example when embedding categorical variables with the aim to fuse them with unstructured data features. In this case, the embedding is trained along the rest of the network, with a mutual influence on the feature extraction process [24, 104].

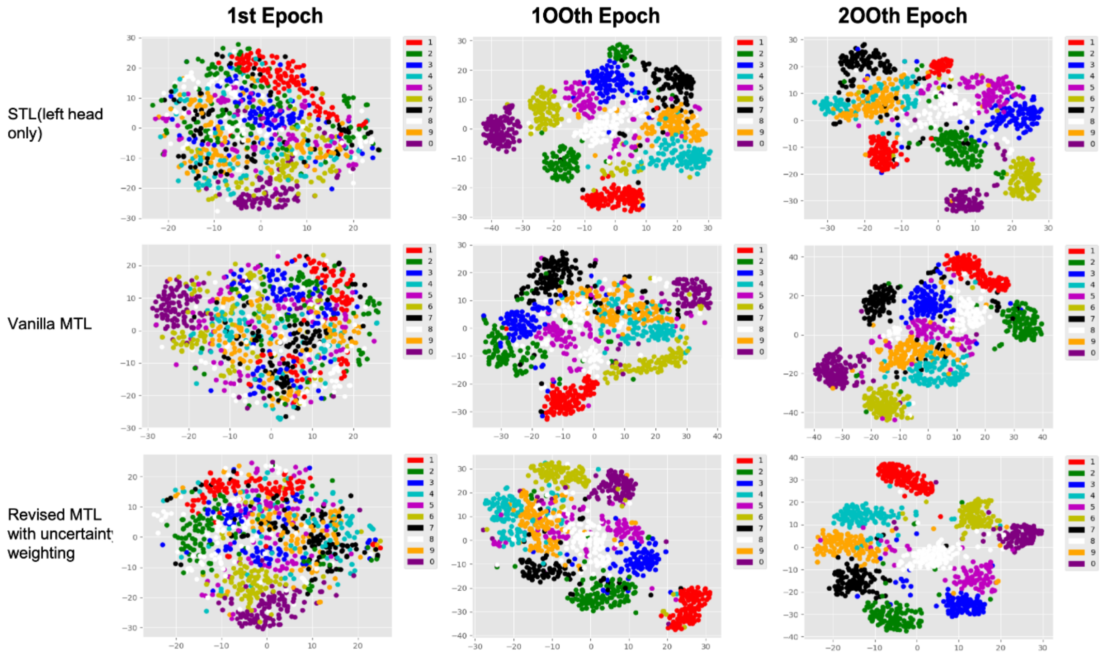


Figure 3.14: Comparison of embeddings learned by ANNs at different stages of their training process (nth epoch). Gong et al. [64], CC BY, no changes.

Chapter 4

Deep Learning in Clinical Dermatology

One of the first modern applications of artificial intelligence (AI) algorithms to medicine was based on experts systems in the 1970s [22]. These algorithms consisted of a knowledge base and an inference engine, which users could query by answering a set of predefined questions. At this period, medical data was scarce, making algorithms based on physician expertise the preferred approaches. This situation gradually evolved with time as data availability and computing power steadily increased. The ensuing progress of machine learning and deep learning enabled researchers to leverage this data and create new applications reshaping medicine in every specialty [53, 158, 193].

Dermatology is a field of medicine particularly suited to AI thanks to the visual accessibility of the skin and the relative facility to acquire image data. The recent breakthroughs in computer vision, especially with convolutional neural networks, have been followed by a myriad of dermatology applications described in several review articles [47, 147, 151, 156]. These applications mostly target dermatologists and general practitioners, but some commercial solutions even aim to advise patients directly [152, 172, 198]. In this chapter, we cover image-based deep learning applications in dermatology, mainly lesion diagnosis and severity grading, highlighting landmarks papers and common technical approaches. Finally, we discuss the opportunities and challenges of the field.

4.1 Deep Learning Applications in Dermatology

4.1.1 Lesion Differential Diagnosis

The majority of studies aimed to support skin cancer screening motivated by the finding that patients' life expectancy strongly improves [177] with the early detection of dangerous forms of cutaneous neoplasia. While cancer screening is an already old research topic [25], the release of public datasets [166] and the organization of challenges[1] by the international skin imaging collaboration (ISIC) in 2016 attracted the attention of the machine learning community as it provided both the data and incentive for researchers to tackle the problem. The best algorithms of the first competition [119] still performed worse than dermatologists, but in 2017, the publication from Esteva et al. [54] was a breakthrough that sparked the interest of both clinicians and researchers. It was the first study of the field to leverage a massive dataset (129'450 clinical and dermoscopy pictures) and achieve performance on par with twenty-one dermatologists for binary classification of benign and malignant lesions. Several publications followed, reporting equivalent or better performance than experts on similar tasks [19, 20, 120, 121]. The top three algorithms of the 2018 ISIC competition outperformed a panel of 511 experts (283 of them board-certified dermatologists) [194] on the diagnosis of seven skin cancer related conditions, illustrating the progress since the first edition of the challenge.

Since the clinical gold standard for skin cancer diagnosis is to perform a biopsy histopathology test, researchers have trained deep learning models (DLMs) to classify histology slides [74, 88, 89, 213]. Heckler et al. [79] even reported better performance than eleven pathologists on a hundred slides for malignancy test, although the publication was received with concerns regarding the study design [60] such as the use of cropped slides.

Researchers have also aimed to perform differential diagnosis beyond skin cancer [224], notably Liu et al. [113] who trained a DLM on 16'114 teledermatology patient cases. Their model was able to diagnose lesion pictures within a list of twenty-six common skin conditions and a larger list counting 419 different diseases. They observed that their DLM was on par with dermatologists but superior to general practitioners and nurses, and concluded that it could perform triage and improve referrals.

The classification of more specific diseases has also been investigated, for example inflammatory skin diseases [212], facial disorders [63] or subtypes of eczema [91] and psoriasis [6, 217]. Chan et al. [27] performed a prospective clinical validation of a classification method for wounds and ulcers, among other publications on the same topic [7, 67, 165]. The diagnosis of onychomycosis was studied based on clinical [73, 96], dermoscopy [225] and histology [43] images.

4.1.2 Lesion Segmentation and Severity Grading

Following the diagnosis of a skin disease, dermatologists evaluate its progression stage by grading its severity. Clinical scoring systems consist in weighting categorical severity scores of disease efflorescences with coarse estimates of lesions' surface, for example using hand surface units [162] (one hand unit corresponds to ~1% of the body surface). There have been three main approaches to automate these systems with deep learning: classification of lesions into severity levels, segmentation of lesions (delineation of their boundaries) to quantify their surface or hybrid classification and segmentation

[1] ISIC Challenge: https://challenge.isic-archive.com (Accessed: 2nd February 2023)

methods.

Automation of the psoriasis PASI index [56] has been widely studied with all three approaches. Schaap et al. [174] trained a classification DLM for each of the PASI features (erythema, desquamation, induration and area) and main body regions (trunk, arms and legs) to predict discrete scores. Meienberger et al. [127] along with other studies [41, 108, 157] segmented psoriasis lesions from patients' pictures and produced precise estimates of their surface. Mooen et al. [136] proposed algorithms for both psoriasis segmentation and severity classification.

Similarly, researchers created classification DLMs for acne severity grading [105, 216]. Seite et al. used a combination of classification and segmentation DLMs to rate acne based on mobile phone pictures [178]. Medela et al. performed a comparable study to grade atopic dermatitis [126] and made their model accessible via a web platform. Other studies targeted skin cancer [110], wounds and ulcers [28, 143, 155], eczema [144], vitiligo [117] and rosacea [15]. Segmented lesions were compared following the ugly duckling concept [68] to find abnormalities [134]. Researchers have also worked on histology slides [89, 202], mainly for skin cancer. Based on lesion segmentation, counts could be inferred and used as markers for severity [125, 176] or to monitor the evolution of diseases.

4.1.3 Common Technical Approaches

The main data sources used in deep learning applications are clinical photographs, dermoscopy images and histology slides. The usual preprocessing procedures involve image resizing, centering, cropping, color calibration and artifacts removal (color scales, clinical markings, etc.). During training, data is usually augmented with random rotations, flips, translations, scaling, color transformations, cutouts, brightness, and contrast changes.

While most published models use transfer learning with ImageNet [44] pretraining, recent works have started to use self-supervised pretraining on datasets from the medical domain instead [14, 29, 134].

Most studies analyze a single image per lesion, but some approaches process several images from the same lesion [113], in which case the image features are first extracted independently and then combined, for example by averaging. When clinical metadata is available, it is either one-hot encoded (unique encoding with sequence of 0 and 1) or embedded before being merged with image features [70, 218].

DLMs in dermatology are usually based on popular convolutional neural network (CNN) architecture such as ResNet [77], Inception [189] or EfficientNet [190] for lesion diagnosis and variations of U-Net [164], Mask R-CNN [76], DeepLab [32] for lesion segmentation. Every winning algorithm [61, 70, 122, 145] of the ISIC competitions was based on ensembles of CNNs, except for the first edition in 2016 [220]. In such configurations, the networks' predictions are usually combined through averaging or voting strategies, although some researchers have also combined them with machine learning algorithms [196].

To guide the training process, researchers usually use variations of the cross-entropy and focal loss for classification with the addition of dice loss for segmentation. The main classification performance metrics are sensitivity (recall), specificity, area under the ROC curve, positive predictive value (precision) and balanced accuracy. For lesion segmentation researchers have used sensitivity, specificity, positive predictive value, intersection over union and dice score.

4.2 Opportunities

While deep learning applications can only tackle restricted tasks, they enable "augmented intelligence" and have the potential to "enhance and scale human expertise" [9, 192]:

- DLMs can assist experts for decision support. A study reported improved skin cancer screening diagnostic accuracy for clinicians assisted by DLMs, especially for general practitioners and dermatologists with less experience [195]. This result also indicates that DLMs can support clinician education.

- Primary care clinicians can be empowered to perform triage [87], improving the accuracy of expert referrals and the adequacy of proposed treatments. The frequency of misdiagnoses and subsequent erroneous treatments could be reduced.

- Teledermatology applications will benefit from an important scale-up by leveraging deep learning (cf. chapter 9). Their synergy could help alleviate the shortage of dermatologists [97, 161], and extend health services to new geographical regions [124, 211].

- Due to their algorithmic nature, DLM predictions are fully automated and reproducible. Repetitive tasks such as mole assessment or histopathology analysis of histology slides can be automated with high precision, thus reducing costs and enabling dermatologists to spend more time with patients.

- Intra- and inter-experts variations [18, 66, 219] can be reduced with DLMs, which generate objective and reproducible quantitative metrics. This enables more precise disease monitoring, benefiting drug development trials and opening the path to personalized medicine.

Overall, if DLMs are implemented as per the recommendations of the American Medical Association [9, 192], the general quality of healthcare services should improve to the advantage of all stakeholders [34].

4.3 Challenges

4.3.1 Comparison with Dermatologists

The fairness of performance comparisons between DLMs and dermatologists is debated, as most studies operate in settings that strongly differ from practice. Researchers have raised concerns on recurring biases [40, 47, 60, 170]. For instance, studies artificially restrict the list of admissible diagnoses, unlike practice where any conditions can arise. Skin cancer screening, for example, is often reduced to a binary malignancy test and other diagnoses are ignored without any considerations on the risks associated to future evolution of the lesions. Furthermore, comparisons are performed on test images, usually from the same source as the training images, which does not reflect the actual generalization capacity of DLMs [45, 194] that is expected from dermatologists.

Dermatologists follow a holistic approach, considering each patient's condition as a whole, while DLMs only analyze a lesion's picture or specific aspects of the patient's condition. Comparing differential diagnosis performance on sole images without giving dermatologists access to patients, their history, or other relevant clinical information differs from what they were trained for and are used to in practice.

Published DLMs are mostly designed to diagnose diseases equivalently, while dermatologists, on the other hand, take particular care to detect diseases with high impact on patients' lives. They can argue in favor of their decisions, while the decision processes of DLMs remain opaque as they are too complex to analyze [206]. Overall, it remains unclear whether DLMs trained from experts' labels can ever be said to outperform dermatologists.

4.3.2 Lack of Data

DLMs can autonomously learn hierarchical features from unstructured data, provided they are trained over sufficiently large datasets. The quality of data is key. Otherwise, training will not converge or predictions will be unreliable [59]. To achieve robust performance, training data should include sufficient samples for every variation (diseases, phenotypes, skin types, picture capture settings, etc.) that could be encountered during inference. This implies that the more general the application range of a model is, the more training data will be required. The extreme case being teledermatology [113].

Conversely, restricting the model application's scope eases these requirements. Imposing protocols on data acquisition [80] by defining constraints on captured body regions, patient posture, zoom level, background also reduces the necessary quantity of data. The use of standardized devices such as full-body imaging systems facilitates this process. For legacy unstandardized datasets, data augmentations can be a mitigation solution [180]. Another possibility is to use generative models to produce synthetic data [2, 58].

Pretraining DLMs with large public datasets such as ImageNet [44] is a popular and effective method to palliate data scarcity [140]. It relies on the assumption that basic features (edges, corners, shapes, etc.) learned by the early convolutional layers will accommodate any vision problem, including dermatology-specific tasks. While clinical data becomes more and more available, it often remains unlabeled due to the lack of resources. In these cases, it is still possible to leverage this data by applying self-supervised pretraining techniques [14, 29, 134].

Data acquisition is also challenging due to the legal and ethical constraints of the medical field. Patients should provide informed consent, and clinical datasets cannot be shared easily between institutions. Consequently, researchers often develop their DLMs based on private monocentric datasets and cannot compare achieved results with their peers. A solution to this problem is to establish public datasets through initiatives like DataDerm [148] and ISIC, or at least, publish anonymized test sets to enable researchers to report comparable performance. When this is not possible, multiple research institutions can collaborate and train DLMs together with federated learning [173]. This method allows to leverage multicentric datasets, improving DLMs robustness, without requiring patient data sharing.

4.3.3 Bias in Data

Any dataset only mirrors facets of reality, and thus introduces biases that need to be mitigated. Since DLMs are trained through the optimization of proxy performance measures, they tend to learn shortcuts rather than solving tasks as expected [54, 209]. Biases can arise from technical artifacts such as clinical markings, device-specific settings, picture background or secondary recurring objects like rulers and color charts. The impact of these artifacts can usually be attenuated using preprocessing techniques such as cropping, manual removal or color calibration.

Most public dermatology datasets contain patients with Fitzpatrick skin type I to IV [3, 4, 103]. This implies that the performance of DLMs trained over such datasets cannot be guaranteed for patients with different skin types [72, 154]. An aggravating factor is that diseases' phenotypes may differ between skin types and introduce additional confusion [38]. Collecting sufficient data from every skin type is thus required.

Another bias is the distribution and development stage of skin diseases. Typical European datasets contain patients afflicted with local diseases [50], usually at an early stage since most patients have the opportunity to receive appropriate care before their state worsens. However, the distributions observed in teledermatology settings or in low-income countries with restricted healthcare systems definitely differ. Diseases' phenotypes also vary depending on the stage of the disease [38]. While these variations can still be recognized by dermatologists, they may be problematic for DLMs. Thus, researchers should validate whether the spectrum of diseases, their development stage and phenotypes are covered appropriately in their training set.

Expert label acquisition is challenging in dermatology context. Many differential diagnosis studies use annotations that do not follow clinical gold standards, which depending on the targeted diseases vary from histopathology biopsy tests to anamnesis-based diagnoses including full skin examination and laboratory tests. Due to practical constraints (costs, retrospective studies, etc.), many studies resort to diagnoses based on sole lesion images, in the best case, by multiple experts from whom a consensus is evaluated [113]. With the inevitable inter-raters variations [18, 66, 219] compounded by the unfamiliarity of most dermatologists with image-based diagnosis, there may be cases where expert's annotations do not match actual clinical diagnoses, introducing noise and raters' bias. Adding to this, Heckler et al. [78] showed, in the context of skin cancer, that DLMs were highly sensitive to label noise. The ideal solution would be to only use data samples with labels produced following clinical gold standards.

In practice, there are also "normal" biases that cannot be fully mitigated. Some diseases have specific predilection sites such as rosacea and onychomycosis, which exclusively affect the head and nails regions respectively. This implies that all pictures from these diseases will inevitably involve these body regions. Depending on the training set, a DLM could learn to recognize these specific body regions rather than the actual diseases. It is important to identify such biases, decide whether they are acceptable, and adopt appropriate verification processes. For rosacea and onychomycosis, a possible mitigating measure would be to include pictures of healthy patients' heads and nails, as well as pictures from other diseases affecting the same regions. To detect biases, it is usually helpful to analyze, which parts of an image have the most influence on DLMs predictions using visualization techniques such as Grad-CAM [179], saliency maps [181] and attention [95]. In general, full automation should be restricted to easily verifiable tasks with controlled negative outcomes. For more complicated cases, it is safer to keep experts in the loop.

4.3.4 Deployment of Deep Learning Models

There are various practical challenges faced during the deployment of DLMs [75], the most obvious being the necessary technical infrastructure with sufficient computational capacity. The complexity and means depend on the context. Launching a deep learning service for a city hospital will entail different challenges than setting-up a teledermatology service in the countryside. Aside from these technical considerations, the translation of DLMs from the research environment to the real world can raise several issues.

We already discussed that training data should reflect the situations where DLMs will be applied and ideally be acquired from the same source. Researchers should also consider that these datasets are fixed in time, whereas the data distribution encountered in practice may change over time [197] (e.g. pandemics). This implies the need to monitor any evolution and update the deployed DLMs accordingly.

Public challenges' winning approaches (e.g. ISIC) are mostly based on ensembles of complex DLMs. These have very high computational requirements and cannot realistically be industrialized for deployment in practice. Furthermore, each of the ensemble's models would need to be maintained and updated over time, which is impractical.

Except for well-studied conditions such as skin cancer, rare diseases are often underrepresented in dermatology datasets or simply ignored due to missing data. This is an issue for most studies' DLMs, as they cannot handle unknown classes and would instead predict the most similar skin condition among the diseases they were trained for. Thus, deployed DLMs should be able to recognize such cases [167] and warn clinicians when needed.

Study design should ensure that both research and practical clinical objectives align. For example, skin cancer screening researchers aim to predict lesion malignancy and usually tune their DLMs to balance sensitivity with specificity. In clinical settings, suspicion of malignancy translates to taking a biopsy for histopathology analysis. When considering the actual need of a biopsy, dermatologists are trained to balance estimated risks with the operation burden for patients. Should these DLMs be deployed in practice, it is unclear whether the chosen sensitivity and specificity compromise could result in an increase in unnecessary biopsies. To prevent such situations, studies should report all metrics relevant for clinical utility, in this case, the precision.

Although it is still a subject of debate [205], the European general data protection regulation (GDPR) requires automated decision to be explainable (art. 22, recital 71) [26]. However, the explainability of DLMs is a well-known problem and open topic of research [183, 201]. With the current American legal framework [199], DLMs are considered as medical devices, implying that they can only aid dermatologists, who keep the full liability of their decisions. Another legal challenge introduced by GDPR is data ownership and privacy. It is required that patients are properly informed on the usage of their data and be able to request both corrections and erasures (art. 12-19). Forgetting data samples is also an open topic of research [114], so it is still unclear how this could be implemented in practice. Furthermore, it was shown that training data could be extracted from deployed DLMs [112], which could put patient data at risk. Data privacy in deep learning and the protection against such attacks is again an open topic of research [17].

4.3.5 Adoption of Deep Learning Models

The advent of new technologies or methods is always met with contrasting opinions, including a certain degree of scepticism and resistance. deep learning applications in dermatology are not an exception.

Some dermatologists are wary that AI imposes a threat to their profession [52, 123], while patients may dread to be treated by machines in the future [142, 223]. Such fears come from generalized misconceptions on the actual capacities of deep learning methods and should fade with better education, starting with dermatologists [139, 151, 175]. Patients should be reassured that dermatologists, on the contrary, will be more available for them as time-consuming tasks are automated [192]. They should understand that automation will reduce inter-experts variations and improve overall care quality. It would be beneficial for research projects to systematically include dermatologists to ensure that clinical utility is among research goals and outcomes. In their review, Zakhem et al. [221] observed that only 41% of pre-February 2019 machine learning publications related to skin cancer included dermatologists among their authors.

Many dermatologists do not believe that the performance of DLMs on research datasets will translate to practice [45, 133]. Freeman et al. [57] analyzed studies on six different skin cancer screening mobile phone applications and concluded that they could not be relied on to detect every melanoma cases and that practice performance is likely poorer than measured in laboratory. This idea is further reinforced by the opacity of DLMs, which are so complex that their decisions cannot be explained [206]. In comparison, dermatologists can explain the rationale behind their choices. To overcome this problem and gain the trust of both clinicians and patients, studies should fulfill the position statement of the American Medical Association [9, 192] on augmented intelligence in health care, which especially advocates that clinical trials must validate actual benefits in practice. Extensions of existing clinical protocols have already been proposed to accommodate for AI-based methods [111, 163]. So far, one of the rare prospective diagnostic accuracy study for melanoma [118] compared the performance of dermatologists with several medical devices including a deep learning based commercial solution on a total of 184 patients whose lesions were excised and assessed by 2 pathologists. The authors concluded that the top device in terms of sensitivity and specificity was the deep learning solution and suggested it could be used to aid clinicians with diagnosis, but would not replace clinical decision-making. While the scope of this finding was contested in a letter [170], which highlighted among other concerns that a sensitivity and specificity analysis was not sufficient to fully assess diagnostic added value, this is already a step in the right direction.

Chapter 5

Anatomy Mapping of Clinical Images of Patients

In this chapter, we propose a method to generate an anatomical region mapping from patients' photographs. The determination of lesions' locations on the body is key for the analysis of skin diseases. However, its automation has not been researched so far. Our mappings can be combined with existing dermatology applications such as lesion detection or segmentation, enabling new kinds of research analysis (e.g. lesion anatomical stratification in section 7.3).

Section 5.1 presents our research article on the macro- and micro-anatomy mapping of patient's photographs. In section 5.2, we extend the micro-anatomy mapping method to other regions of the body.

5.1 Automated Anatomical Mapping of Skin Photographs

This research article was published [11] at the journal of the European Academy of Dermatology and Venereology [1]. Our work was based on two main hypotheses. First, we hypothesized that a deep learning model could be trained to recognize anatomy regions in patches of skin photographs with at least the same performance as humans. Second, we hypothesized that including the lesion location in the training of an image-based differential diagnosis classifier would benefit performance.

Both hypotheses were confirmed: we propose an approach to perform coarse localisation of the main body regions and observe improved differential diagnosis performance when leveraging this information together with lesion images. Furthermore, we suggest a method to produce fine anatomical mapping of the ear region with sufficient precision to assist dermatologists in lesion documentation. The location information is one of the first features of lesion dermatological description and has an influence on differential diagnosis since skin diseases may have predilection sites. Its determination is thus an important preliminary step in the analysis of skin lesions.

[1] Full text via DOI: `https://doi.org/10.1111/jdv.18476` (Accessed: 2nd February 2023)

DOI: 10.1111/jdv.18476

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 dermatosurgical 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

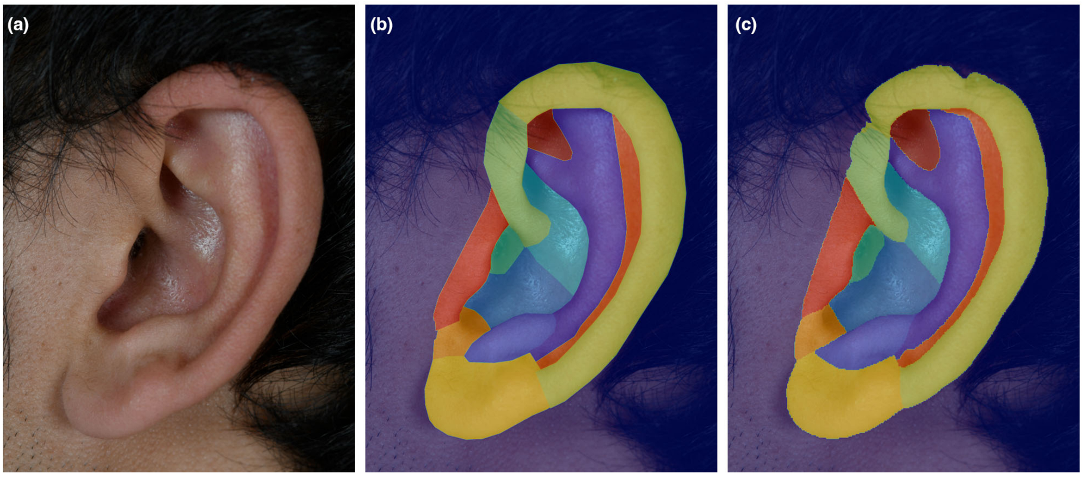
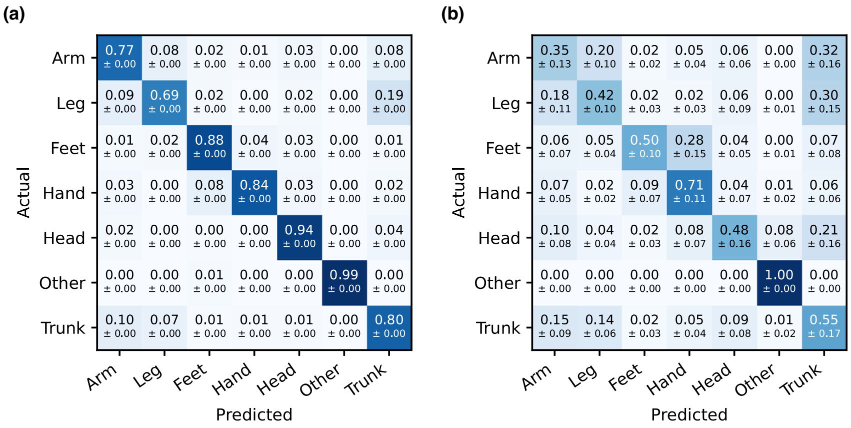


Figure 1 Ear test sample (a) with expert's annotations (b) and DLM's predictions (c). Picture randomly selected from the test set. The original image is shown in (a), the expert's annotation in (b) and the DLM's predictions in (c). The regions are coloured as follows: anti-helix in violet, anti-tragus 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.



(a)

Actual \ Predicted	Arm	Leg	Feet	Hand	Head	Other	Trunk
Arm	0.77 ± 0.00	0.08 ± 0.00	0.02 ± 0.00	0.01 ± 0.00	0.03 ± 0.00	0.00 ± 0.00	0.08 ± 0.00
Leg	0.09 ± 0.00	0.69 ± 0.00	0.02 ± 0.00	0.00 ± 0.00	0.02 ± 0.00	0.00 ± 0.00	0.19 ± 0.00
Feet	0.01 ± 0.00	0.02 ± 0.00	0.88 ± 0.00	0.04 ± 0.00	0.03 ± 0.00	0.00 ± 0.00	0.01 ± 0.00
Hand	0.03 ± 0.00	0.00 ± 0.00	0.08 ± 0.00	0.84 ± 0.00	0.03 ± 0.00	0.00 ± 0.00	0.02 ± 0.00
Head	0.02 ± 0.00	0.00 ± 0.00	0.00 ± 0.00	0.00 ± 0.00	0.94 ± 0.00	0.00 ± 0.00	0.04 ± 0.00
Other	0.00 ± 0.00	0.00 ± 0.00	0.01 ± 0.00	0.00 ± 0.00	0.00 ± 0.00	0.99 ± 0.00	0.00 ± 0.00
Trunk	0.10 ± 0.00	0.07 ± 0.00	0.01 ± 0.00	0.01 ± 0.00	0.01 ± 0.00	0.00 ± 0.00	0.80 ± 0.00

(b)

Actual \ Predicted	Arm	Leg	Feet	Hand	Head	Other	Trunk
Arm	0.35 ± 0.13	0.20 ± 0.10	0.02 ± 0.02	0.05 ± 0.04	0.06 ± 0.06	0.00 ± 0.00	0.32 ± 0.16
Leg	0.18 ± 0.11	0.42 ± 0.10	0.02 ± 0.03	0.02 ± 0.03	0.06 ± 0.09	0.00 ± 0.01	0.30 ± 0.15
Feet	0.06 ± 0.07	0.05 ± 0.04	0.50 ± 0.10	0.28 ± 0.15	0.04 ± 0.05	0.00 ± 0.01	0.07 ± 0.08
Hand	0.07 ± 0.05	0.02 ± 0.02	0.09 ± 0.07	0.71 ± 0.11	0.04 ± 0.07	0.01 ± 0.02	0.06 ± 0.06
Head	0.10 ± 0.08	0.04 ± 0.04	0.02 ± 0.03	0.08 ± 0.07	0.48 ± 0.16	0.08 ± 0.06	0.21 ± 0.16
Other	0.00 ± 0.00	0.00 ± 0.00	0.00 ± 0.00	0.00 ± 0.00	0.00 ± 0.00	1.00 ± 0.00	0.00 ± 0.00
Trunk	0.15 ± 0.09	0.14 ± 0.06	0.02 ± 0.03	0.05 ± 0.04	0.09 ± 0.08	0.01 ± 0.02	0.55 ± 0.17

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

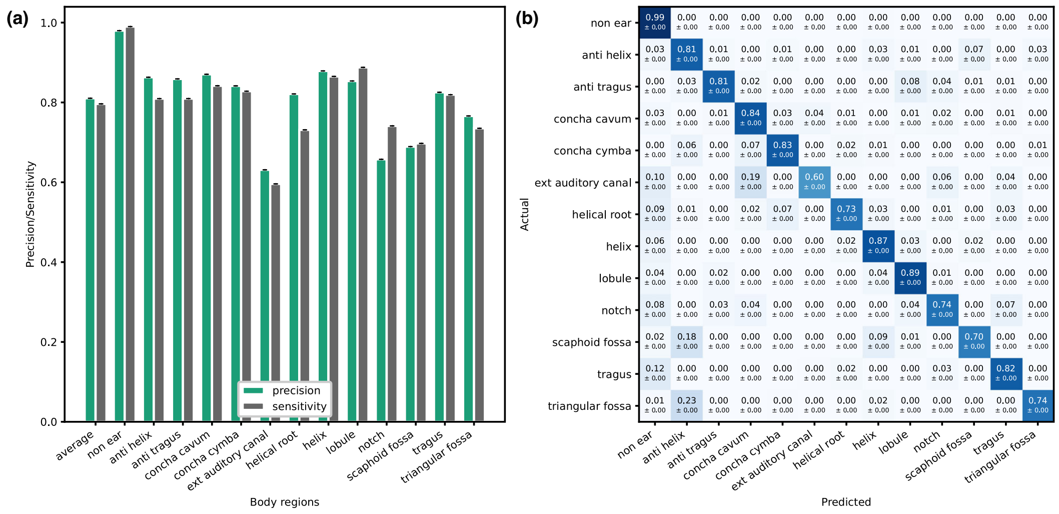


Figure 3 Ear segmentation DLM's micro-anatomical performance. The values show the proportion of images ± SD. (a) Precision and sensitivity: the average pixel precision and sensitivity reached on the test set by the DLM. (b) Confusion matrix: the average pixel proportion segmented among the 12 ear regions 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.

SUPPORTING INFORMATION

Performance Metrics

Precision and sensitivity are metrics used to evaluate the performance of deep learning models (DLM). For a given body region X, precision corresponds to the fraction of valid predictions for location X, while sensitivity is the fraction of patches from location X that were correctly detected by the model. They are computed as follows:

$$Precision = \frac{Truepositive}{Truepositive + Falsepositive}$$

$$Sensitivity = \frac{Truepositive}{Truepositive + Falsenegative}$$

Body regions separation

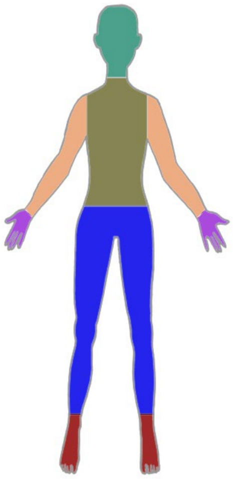


Figure S1 Body regions

Example of skin patches with increasing localization difficulty

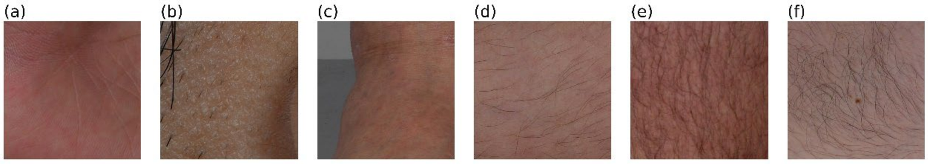


Figure S2 Examples of skin patches with increasing localization difficulty.
From left to right, the patches come from the following body regions: hands (a), head (b), feet (c), arms (d), legs (e) and trunk (f).

Table S1 Body regions dataset patch distribution

	Arm	Leg	Foot	Hand	Head	Trunk	Other	Total
Weak labels	40029	43212	45046	32375	61987	24495	29978	277122
Strong labels (train)	4155	2656	3374	3585	3879	3467	2999	24115
Strong labels (test)	510	510	510	510	510	510	510	3570

The weak labels' distribution is included for completeness but varied during the training process, as the labels were continuously re-evaluated.

Table S2 Macro-anatomy experts performance

	Dermatologists		Students	
Region	Precision	Sensitivity	Precision	Sensitivity
Arm	47% (33-60)	29% (20-36)	42% (23-86)	38% (11-55)
Leg	49% (42-61)	43% (28-59)	49% (34-65)	42% (25-52)
Feet	82% (50-99)	53% (28-63)	76% (58-93)	49% (40-66)
Hand	70% (43-83)	75% (65-87)	58% (49-70)	69% (47-89)
Head	67% (39-91)	49% (29-80)	68% (49-88)	47% (29-68)
Other	90% (81-100)	100% (100-100)	91% (80-99)	100% (100-100)
Trunk	39% (29-48)	65% (45-83)	40% (27-62)	49% (21-73)
Average	63% (58-68)	59% (53-63)	61% (56-70)	56% (52-65)

There were 6 dermatologists and 12 medical students. The value in parentheses correspond to the 95% confidence interval.

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

Region	Precision	Sensitivity
Arm	76% (72-80)	73% (69-76)
Leg	75% (72-79)	69% (65-73)
Feet	86% (83-89)	79% (75-82)
Hand	88% (86-91)	80% (76-83)
Head	80% (76-82)	90% (87-92)
Other	98% (97-99)	98% (97-99)
Trunk	66% (62-69)	78% (74-81)
Average	81% (80-83)	81% (80-82)

The model was trained using strongly labeled data only, without any weakly labeled data. The values in parentheses correspond to the 95% confidence interval.

Generation of macro-anatomical body regions mapping

Image randomly selected from the test set. The patches are drawn as white dashed squares and their corresponding top two predictions are written in yellow. The model predicted the localization of each of these patches independently without using the rest of the image. Although some patch predictions are wrong, the main body regions, in this picture the legs and feet, can be correctly determined.
The first label corresponds to the most probable location predicted by the DLM. The second corresponds to the second most probable among the remaining regions.

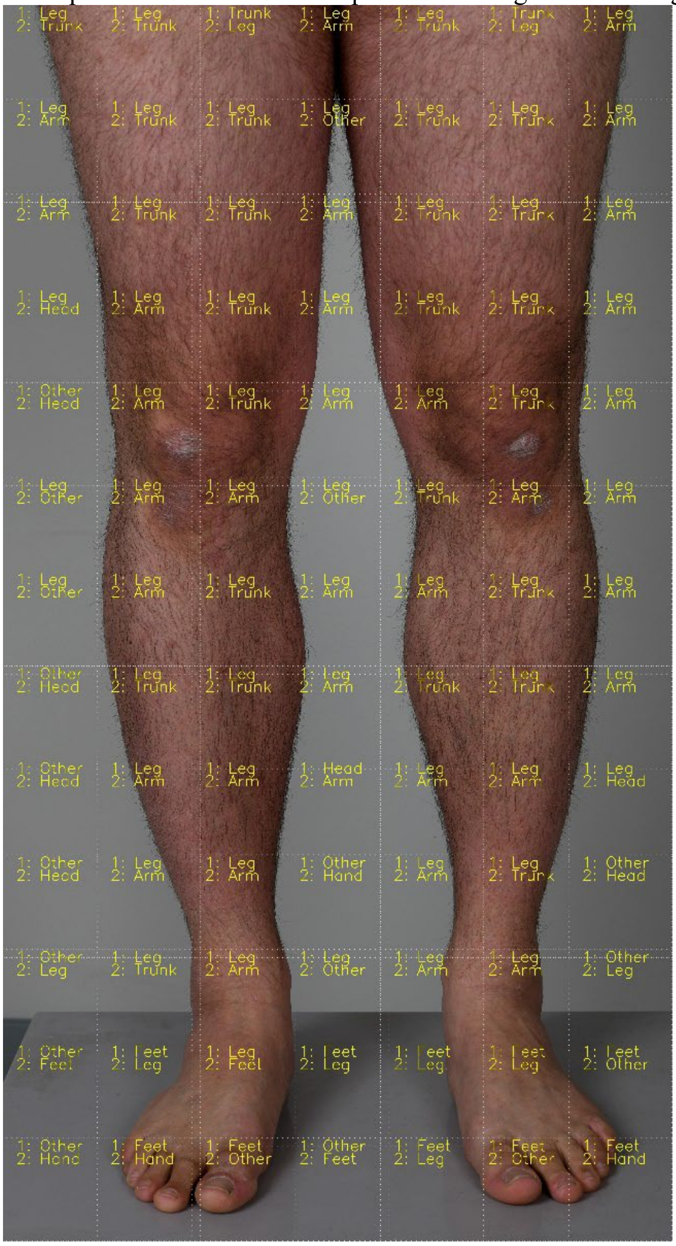


Figure S3 Generation of macro-anatomical body regions mapping

Differential diagnosis with/without lesion localization metadata

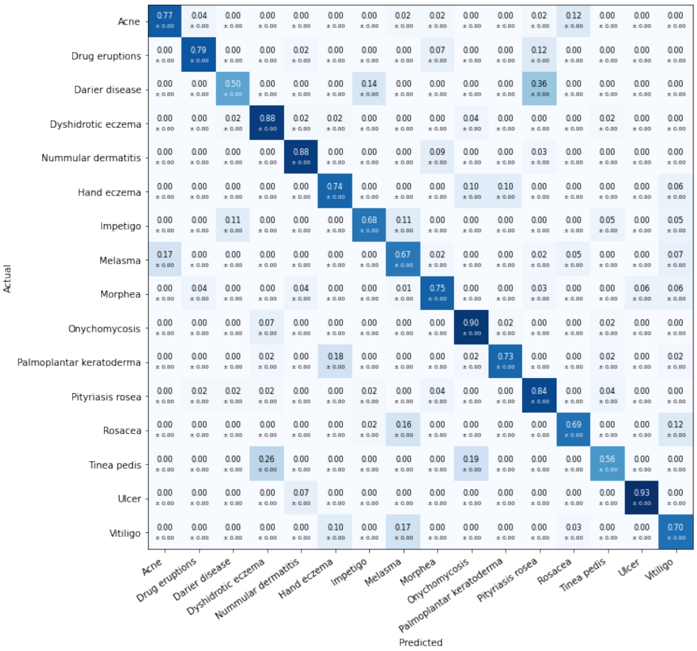


Figure S4 Confusion matrix of DLM trained with lesion's pictures only. The values show the proportion of images +/- SD.

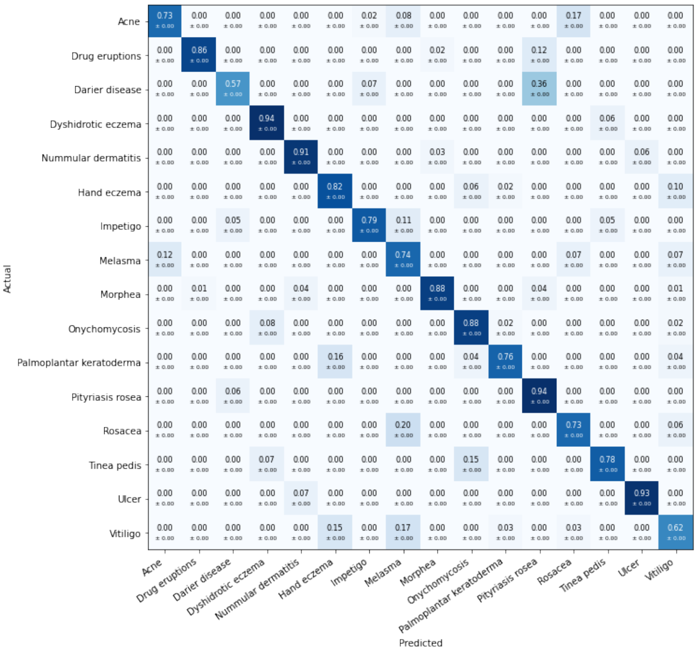


Figure S5 Confusion matrix of DLM trained with both lesion's locations and pictures. The values show the proportion of images +/- SD.

Table S4 Differential diagnosis specificity

Diseases	Test images	Specificity A	Specificity B
Acne	48	99% (98-100)	99% (98-100)
Drug eruptions	43	99% (98-100)	100% (100-100)
Darier disease	14	99% (99-100)	99% (99-100)
Dyshidrotic eczema	50	98% (97-99)	99% (98-100)
Nummular dermatitis	34	99% (98-100)	99% (98-100)
Hand eczema	50	98% (97-99)	98% (97-99)
Impetigo	19	99% (99-100)	100% (99-100)
Melasma	42	97% (96-98)	96% (95-98)
Morphea	68	98% (97-99)	100% (99-100)
Onychomycosis	60	98% (97-99)	99% (98-99)
Palmoplantar keratoderma	45	99% (98-100)	100% (99-100)
Pityriasis rosea	50	98% (96-99)	98% (97-99)
Rosacea	49	99% (98-99)	98% (97-99)
Tinea pedis	27	99% (98-100)	99% (99-100)
Ulcer	41	99% (99-100)	100% (99-100)
Vitiligo	40	97% (96-98)	98% (97-99)
Average	-	98% (98-99)	99% (99-99)

Specificity 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.

Table S5 Predilection sites of the different diagnoses

Diagnoses	Train images	Test images	Predilection sites
Acne	189	48	Head, trunk
Drug eruptions	168	43	Trunk, arms, legs
Darier disease	55	14	Trunk, head
Dyshidrotic eczema	195	50	Feet, hands
Nummular dermatitis	135	34	Legs, trunk, head, arms
Hand eczema	196	50	Hands
Impetigo	74	19	Head, trunk, arms
Melasma	165	42	Head
Morphea	271	68	Trunk, arms, legs
Onychomycosis	231	60	Feet, hands
Palmoplantar keratoderma	177	45	Feet, hands
Pityriasis rosea	195	50	Trunk
Rosacea	192	49	Head, trunk
Tinea pedis	104	27	Feet
Ulcer	162	41	Legs
Vitiligo	158	40	Hands, feet, head, trunk

5.2 Micro-Anatomical Region Mapping of the Human Body

With the ambition to map the whole human body, we extended the micro-anatomy approach to other anatomical regions. In this section, we present the results achieved so far. Except for the ear and hand anatomy models, these results were not published yet.

5.2.1 Materials and Methods

Data We created datasets for the different body regions (cf. table 5.1) and organized the labeling process with medical students under the supervision of a board-certified dermatologist specialist in the human anatomy. The test sets were generated by randomly sampling 20% of the images, ensuring no leaks of pictures from the same patient.

Body Region	**# Images**	**# Sub-regions**	**Status**
Eye	129	10	Completed
Ear	182	13	Completed
Mouth	131	9	Completed
Nose	215	14	Labeling
Head	220	19	Labeling
Hand	215	37	Completed
Nail	200	11	Completed

Table 5.1: Overview of micro-anatomy datasets.

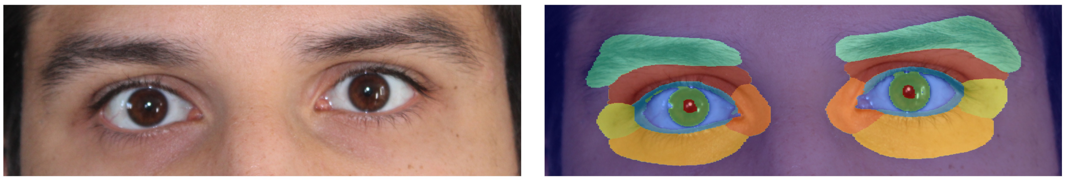
Models For each dataset, we trained separate U-Nets [164] with ResNet50 [77] backbones, and an input size of 380 pixels following the micro-anatomy segmentation approach described in section 5.1. To reduce computation costs and with the aim to directly generate anatomical mappings from full-body images, we also tested training a single deep learning model (DLM) on all datasets grouped together.

5.2.2 Results

We present the performance achieved for the different body regions (tables 5.2 to 5.6) together with sample predictions (figures 5.1 to 5.5).

	Eye DLM		All regions DLM	
Region	**Precision**	**Sensitivity**	**Precision**	**Sensitivity**
Non-eye	83% (76-89)	93% (89-95)	82% (74-89)	91% (87-94)
Conjunctiva	94% (91-95)	90% (85-92)	93% (91-95)	89% (84-92)
Eye margin (lashes root)	67% (62-73)	87% (84-90)	70% (64-75)	84% (81-86)
Eyebrow	84% (77-89)	75% (59-85)	85% (79-89)	73% (57-83)
Iris	92% (85-95)	95% (93-96)	91% (83-95)	94% (92-96)
Lateral canthus	74% (62-84)	60% (51-69)	59% (47-72)	67% (57-74)
Lower eyelid	83% (74-90)	81% (77-84)	80% (71-86)	78% (71-84)
Medial canthus	71% (63-78)	73% (63-83)	71% (64-79)	67% (55-76)
Pupil	89% (85-91)	96% (94-98)	90% (86-93)	97% (95-98)
Upper eyelid	90% (86-92)	75% (69-82)	93% (91-96)	66% (57-73)
Average	83% (80-85)	82% (80-85)	81% (78-84)	81% (77-83)

Table 5.2: Eye anatomy DLMs performance.



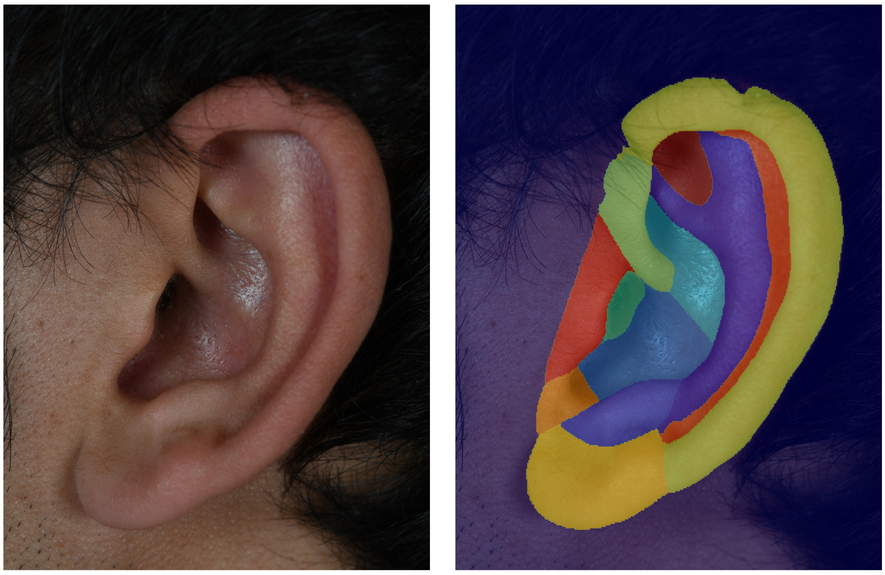
(a) Patient's image

(b) Eye DLM's predictions

Figure 5.1: Micro anatomy of the eye.

	Ear DLM		All regions DLM	
Region	**Precision**	**Sensitivity**	**Precision**	**Sensitivity**
Non-ear	98% (97-99)	99% (99-99)	98% (98-99)	98% (98-99)
Antihelix	86% (82-91)	81% (76-85)	87% (83-92)	79% (74-84)
Antitragus	86% (81-90)	80% (75-85)	85% (80-89)	81% (78-84)
Concha cavum	87% (84-90)	82% (75-89)	88% (84-91)	82% (75-89)
Concha cymba	85% (82-89)	79% (71-87)	82% (78-87)	81% (71-91)
External auditory canal	62% (52-71)	61% (47-75)	62% (52-72)	64% (50-79)
Helical root	80% (77-84)	73% (66-81)	79% (75-83)	72% (66-81)
Helix	88% (84-91)	87% (83-89)	89% (87-92)	85% (83-88)
Lobule	85% (81-88)	90% (86-93)	88% (84-91)	88% (85-91)
Notch	66% (58-72)	76% (71-82)	69% (64-75)	72% (67-77)
Scaphoid fossa	68% (60-74)	70% (65-77)	65% (58-70)	75% (71-80)
Tragus	82% (79-85)	83% (80-86)	81% (78-85)	84% (79-87)
Triangular fossa	73% (65-84)	76% (71-81)	68% (59-82)	79% (76-84)
Average	80% (78-82)	80% (76-83)	80% (78-82)	80% (77-84)

Table 5.3: Ear anatomy DLMs performance.

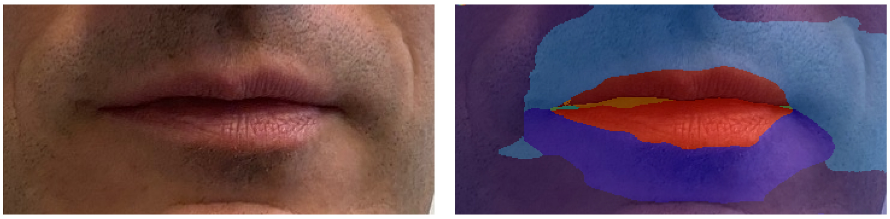


(a) Patient's image (b) Ear DLM's predictions

Figure 5.2: Micro anatomy of the ear.

	Mouth DLM		**All regions DLM**	
Region	**Precision**	**Sensitivity**	**Precision**	**Sensitivity**
Non-mouth	80% (64-91)	62% (50-70)	62% (45-76)	67% (51-76)
Lower lip	73% (62-79)	74% (61-84)	76% (64-85)	66% (51-75)
Upper lip	80% (70-86)	75% (64-82)	90% (85-94)	62% (50-71)
Inside mouth	81% (71-88)	82% (71-90)	74% (61-82)	86% (77-92)
Oral commissure	13% (8-21)	15% (9-25)	14% (10-18)	7% (3-17)
Teeth	96% (93-98)	85% (79-92)	95% (90-97)	77% (67-87)
Tongue	75% (46-91)	90% (85-96)	77% (50-88)	88% (81-94)
Lower vermilion	65% (51-77)	81% (75-88)	65% (49-79)	85% (80-89)
Upper vermilion	65% (57-74)	63% (51-77)	67% (56-79)	62% (49-74)
Average	70% (65-73)	70% (65-74)	69% (63-73)	67% (62-71)

Table 5.4: Mouth anatomy DLMs performance.

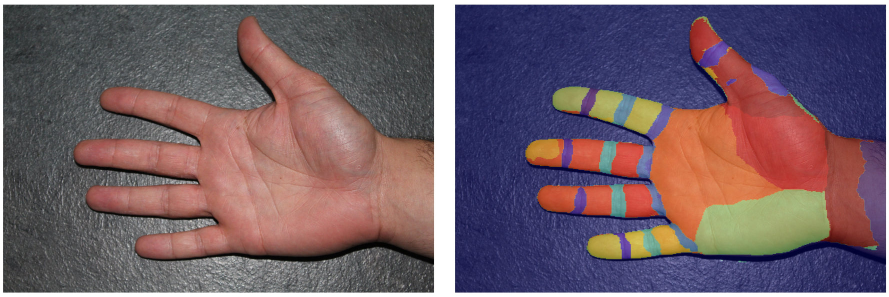


(a) Patient's image (b) Mouth DLM's predictions

Figure 5.3: Micro anatomy of the mouth.

	Hand DLM		**All regions DLM**	
Region	**Precision**	**Sensitivity**	**Precision**	**Sensitivity**
Non-hand	99% (99-99)	97% (97-98)	99% (99-99)	98% (97-98)
DIP2	71% (58-79)	82% (72-88)	75% (66-82)	82% (74-89)
DIP3	77% (72-81)	84% (74-90)	78% (73-82)	85% (79-89)
DIP4	72% (67-78)	84% (73-90)	74% (68-81)	82% (72-89)
DIP5	75% (69-80)	85% (80-90)	77% (70-82)	85% (81-89)
IP	79% (76-82)	84% (81-87)	82% (78-84)	84% (81-86)
MCP1	64% (57-71)	79% (74-84)	66% (58-72)	78% (73-83)
MCP2	74% (69-79)	82% (74-86)	78% (75-82)	83% (79-86)
MCP3	75% (69-79)	84% (79-88)	74% (65-81)	83% (79-86)
MCP4	68% (60-75)	77% (69-83)	71% (65-76)	76% (70-81)
MCP5	72% (65-77)	79% (75-84)	74% (69-78)	81% (78-85)
PIP2	84% (75-90)	88% (82-92)	85% (77-90)	89% (86-92)
PIP3	87% (84-90)	85% (72-91)	85% (80-89)	86% (73-92)
PIP4	84% (78-88)	87% (84-90)	86% (84-88)	88% (83-91)
PIP5	84% (79-87)	86% (82-89)	85% (80-88)	87% (83-91)
Dorsal mid	72% (67-77)	76% (69-81)	76% (72-80)	73% (63-80)
Dorsal radial	86% (81-89)	85% (82-88)	88% (84-91)	86% (84-89)
Dorsal ulnar	87% (85-89)	77% (69-81)	87% (84-89)	81% (77-84)
Hypothenar	87% (84-90)	89% (81-95)	88% (86-91)	89% (83-94)
Index distal	85% (78-92)	88% (82-92)	86% (80-92)	88% (83-93)
Index middle	84% (74-91)	88% (83-92)	89% (84-92)	87% (82-91)
Index proximal	87% (81-92)	89% (83-93)	87% (82-92)	91% (88-93)
Little f. distal	90% (87-93)	89% (82-93)	91% (89-94)	90% (84-94)
Little f. middle	91% (89-93)	85% (82-88)	90% (86-93)	86% (83-89)
Little f. proximal	87% (85-90)	88% (86-91)	87% (85-90)	91% (89-92)
Middle f. distal	91% (87-94)	87% (80-93)	91% (87-94)	88% (80-93)
Middle f. middle	92% (87-94)	88% (82-92)	89% (81-94)	91% (88-93)
Middle f. proximal	89% (85-92)	88% (80-92)	86% (79-91)	90% (82-94)
Nail	89% (86-91)	83% (78-86)	90% (88-92)	83% (78-87)
Palm	89% (86-93)	86% (84-89)	89% (86-92)	88% (84-91)
Ring f. distal	87% (82-92)	87% (76-93)	87% (81-92)	88% (80-95)
Ring f. middle	89% (83-94)	86% (78-91)	90% (85-94)	86% (76-91)
Ring f. proximal	88% (84-91)	88% (84-91)	88% (85-90)	89% (85-91)
Thenar	88% (83-91)	89% (85-92)	90% (85-93)	88% (84-91)
Thumb distal	92% (90-93)	89% (86-92)	92% (90-94)	91% (89-93)
Thumb proximal	87% (83-90)	80% (76-83)	88% (85-89)	82% (76-85)
Wrist	69% (64-74)	86% (83-89)	70% (66-76)	86% (82-90)
Average	83% (80-85)	85% (82-88)	84% (82-86)	86% (83-88)

Table 5.5: Hand anatomy DLMs performance.

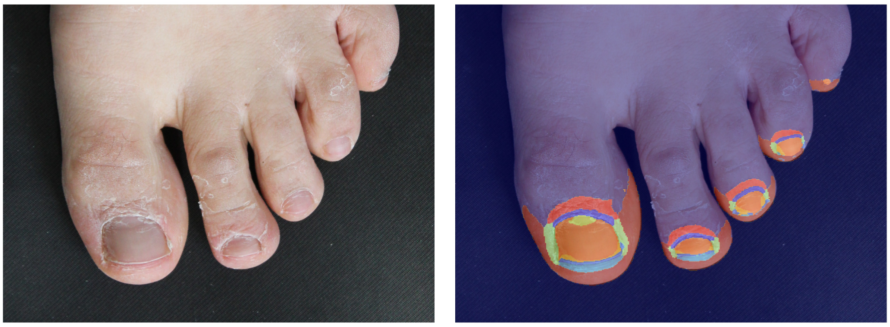


(a) Patient's image (b) Hand DLM's predictions

Figure 5.4: Micro anatomy of the hand.

	Nail DLM		All regions DLM	
Region	**Precision**	**Sensitivity**	**Precision**	**Sensitivity**
Non-nail	99% (99-100)	98% (98-99)	99% (99-99)	93% (90-96)
Cuticle	65% (59-69)	72% (69-76)	62% (56-67)	68% (65-72)
Distal edge plate	60% (37-79)	77% (70-82)	57% (37-74)	73% (67-79)
Distal groove	57% (51-63)	59% (47-77)	54% (47-60)	58% (49-74)
Hyponychium	18% (0-47)	60% (0-65)	18% (0-100)	0% (0-0)
Lateral fold	67% (62-72)	80% (77-83)	63% (58-67)	78% (74-82)
Lunula	81% (65-89)	82% (74-90)	78% (62-87)	81% (72-88)
Onychodermal band	38% (28-48)	48% (37-62)	31% (20-42)	50% (41-62)
Plate	90% (86-93)	81% (73-86)	90% (87-94)	78% (69-84)
Proximal fold	64% (57-71)	75% (70-80)	63% (56-69)	72% (66-78)
Pulp	83% (79-87)	85% (81-88)	81% (76-84)	80% (77-84)
Average	66% (62-69)	74% (67-77)	63% (59-70)	67% (64-69)

Table 5.6: Nail anatomy DLMs performance.



(a) Patient's image (b) Nail DLM's predictions

Figure 5.5: Micro anatomy of the nail.

5.2.3 Discussion

The performance of the DLM trained on all regions is remarkably close to the region-specific DLMs. The main challenging regions were the relatively smaller ones such as the external auditory canal for the ear, the oral commissure for the mouth and the hyponychium for the nails. These regions have the particularity that they can be hidden depending on the position of the patient. Their small size also results in a high pixel imbalance in favor of the other regions, which prevents effective learning despite the customized loss function guiding the training process. The annotation of the images was also a challenge as the boundaries of the anatomical regions are not well-defined theoretically and can vary in the literature. Unclear cases were validated with a board-certified dermatologist, specialist of the human anatomy.

As we already discussed in section 5.1, our approach can assist with lesion description and differential diagnosis. Other practical applications include supporting medical education and enabling targeted image retrieval of specific anatomical regions in dermatology databases. Our method can also be combined with disease segmentation DLMs to determine lesions' anatomical stratification and refine clinical severity grading systems (cf. section 7.3).

Chapter 6

Differential Diagnosis of Skin Lesion Images

6.1 Introduction

The usual differential diagnosis process followed by dermatologists is composed of several steps including the study of patient history, the dermatological description of lesions and, if necessary, additional laboratory tests (cf. section 2.2). Based on the information determined at each stage, dermatologists follow decision trees to determine the differential diagnosis of the disease. When only lesion images are available (e.g. teledermatology context), the usable information for differential diagnosis is reduced to a partial dermatological description since features such as temperature or consistency cannot be assessed.

In similar context, researchers have trained deep learning models (DLMs) to classify lesion images and automate differential diagnosis (cf. section 4.1.1). These approaches are based solely on image statistical features (autonomously extracted by the DLMs) and do not leverage any aspect of dermatologists' differential diagnosis process. Furthermore, it is not possible to interpret these features (cf. section 4.3 and the lack of explainability of DLMs [183, 201]), which are very different from the features used by dermatologists. The consequence is that DLMs' predictions cannot be justified in dermatologists' terms, creating suspicion and precluding adoption.

We propose a hybrid approach to mitigate this issue by letting DLMs leverage both image statistical features and some dermatological description features that can be inferred from a lesion's image. To the best of our knowledge, this direction was not explored in the scientific literature, where models are trained based on lesion images only, eventually with patient metadata (cf. section 4.1.1). Section 5.1 showed that combining lesion location with images features could benefit performance. Here, we hypothesized that the inclusion of efflorescence information would also lead to improved performance. The following sections describe an approach to train a DLM with both lesion image and efflorescence information, together with the achieved performance. These results were not published yet.

6.2 Materials and Methods

Data In this retrospective study, a dataset composed of 1098 hand pictures of patients either healthy or afflicted by eczema, lentigo, psoriasis or vitiligo was labeled by a student for visible efflorescences. The dataset distribution is shown in figure 6.1. Picture diagnoses were obtained from the hospital database. All patients had skin type 1 to 3 on the Fitzpatrick scale. 20% of the images were selected for the test set using stratified sampling on the diagnoses. In addition, 100 images similarly sampled from the test set were labeled by three dermatologists both for diagnosis and efflorescences.

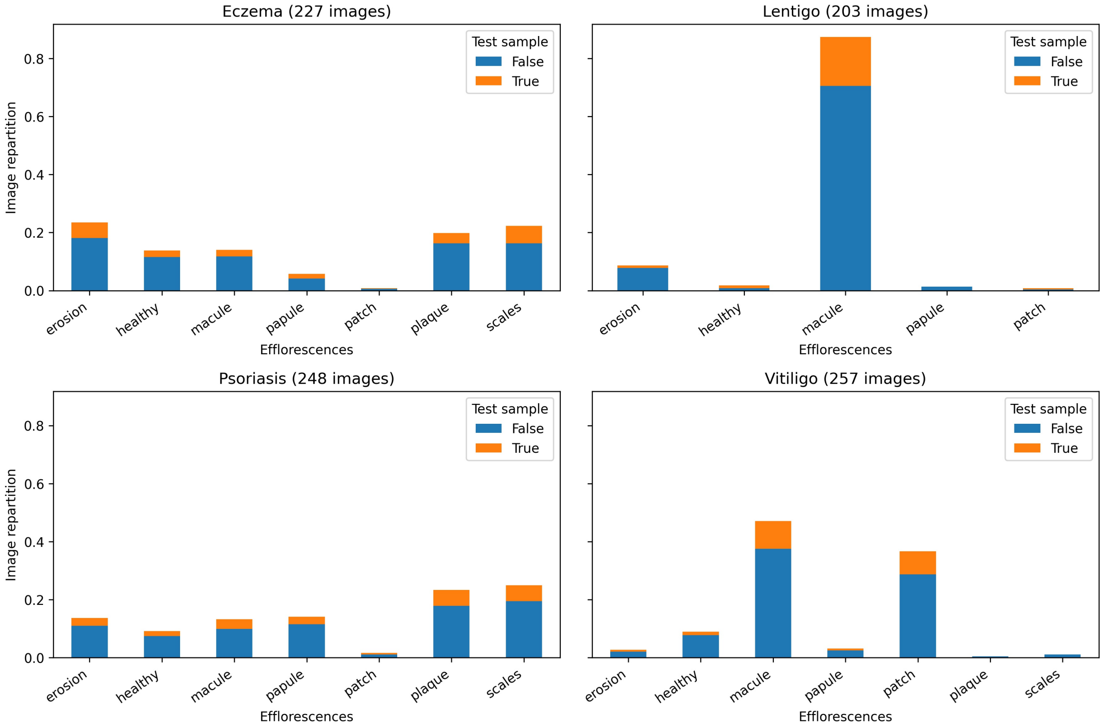


Figure 6.1: Efflorescence dataset distribution. The dataset includes in addition 163 healthy pictures without any efflorescences.

Models Four different differential diagnosis models were trained to assess the benefit of using lesion efflorescence information:

(M_1) Gradient boosting classification algorithm trained only with the efflorescence labels.

(M_2) ResNet18 [77] trained only with the lesion images.

(M_3) ResNet18 trained with both lesion images and efflorescences. The efflorescences were one-hot encoded and appended to the image features extracted by the model backbone.

(M_4) ResNet18 trained with both lesion images and embedded efflorescences. The model learned a 64 dimensional embedding for efflorescences, which was appended to the image features extracted by the model backbone.

6.3 Results

The models' precision and sensitivity with bootstrapped 95% confidence interval are shown in table 6.1, while the confusion matrices are presented in figure 6.2. The most confused classes are eczema with psoriasis followed by vitiligo and then lentigo, which is very well identified.

As could be expected, the healthy class is perfectly identified by M_1. However, the convolutional neural networks achieve this result only when the efflorescence information is embedded. This is probably caused by the large dimension difference between image features and the one-hot encoded efflorescences. While M_3 surpasses M_2, there is no clear benefit in training using both lesion image and efflorescence when compared with the performance achieved by M_1.

The best performing DLM is M_4. To compare this model with the others, we applied the McNemar's test on the respective test set predictions and could confirm a significant difference in error proportions with p-value below 0.05 in every case. Furthermore, we tested the randomness of our results on twenty different initialization seeds (summary available in table 6.2) and observed a systematic superiority of M_4 over the other models.

To put these results in perspective, we evaluated the dermatologists' differential diagnosis performance on a sample of 100 test images and measured an average precision of 76% and sensitivity of 74.6%.

6.4 Discussion

The results illustrate that leveraging both image lesion and efflorescence information can benefit differential diagnosis DLM performance. Taken together with a similar observation for lesion location (cf. section 5.1), this indicates that the hybrid approach of combining statistical features with traditional differential diagnosis features has potential to improve DLM performance and should be further tested, starting with other dermatological description features.

Dermatological features are easier to infer than differential diagnosis since their determination is part of standard lesion assessment. They can be provided to the system directly by the dermatologist or be predicted either in multitask settings or by a separate DLM. Dermatologists can then validate these intermediate predictions, eventually correct them if needed, and observe the impact on the resulting differential diagnosis. Thus, besides the performance gain, our hybrid approach enables dermatologists to interact and better understand aspects of the DLM decision process. Theoretically, this

	M_1		M_2	
Disease	**Precision**	**Sensitivity**	**Precision**	**Sensitivity**
Eczema	57% (44-70)	49% (38-65)	55% (40-67)	51% (36-62)
Healthy	100% (100-100)	100% (100-100)	52% (32-69)	48% (31-66)
Lentigo	58% (48-70)	95% (90-100)	76% (65-89)	98% (92-100)
Psoriasis	60% (45-77)	59% (50-71)	62% (46-74)	67% (52-79)
Vitiligo	95% (89-100)	65% (54-78)	82% (69-90)	67% (54-77)
Macro avg	74% (71-79)	74% (70-79)	65% (59-71)	66% (60-72)
Weighted avg	73% (70-79)	70% (66-76)	67% (61-72)	67% (59-72)

	M_3		M_4	
Disease	**Precision**	**Sensitivity**	**Precision**	**Sensitivity**
Eczema	62% (48-77)	60% (47-74)	72% (56-81)	72% (59-86)
Healthy	66% (53-80)	79% (68-95)	100% (100-100)	97% (88-100)
Lentigo	78% (69-90)	98% (93-100)	84% (73-93)	95% (90-100)
Psoriasis	72% (55-84)	67% (52-80)	72% (60-84)	75% (63-84)
Vitiligo	89% (77-96)	73% (59-83)	88% (79-95)	78% (66-88)
Macro avg	73% (67-79)	75% (69-81)	83% (79-87)	83% (79-88)
Weighted avg	74% (67-81)	74% (67-80)	82% (77-87)	82% (77-86)

Table 6.1: Performance of the differential diagnosis models.

Model	**Precision mean**	**Precision std**	**Sensitivity mean**	**Sensitivity std**
M_1	74.1%	0.00%	73.6%	0.00%
M_2	60.0%	2.24%	59.9%	2.04%
M_3	74.4%	2.56%	73.4%	2.62%
M_4	82.3%	1.44%	82.3%	1.39%

Table 6.2: Evaluation of the performance randomness.

idea could be extended to a fully automated and explainable system, which would predict every feature of the differential diagnosis process and combine them following the same decision trees used by dermatologists.

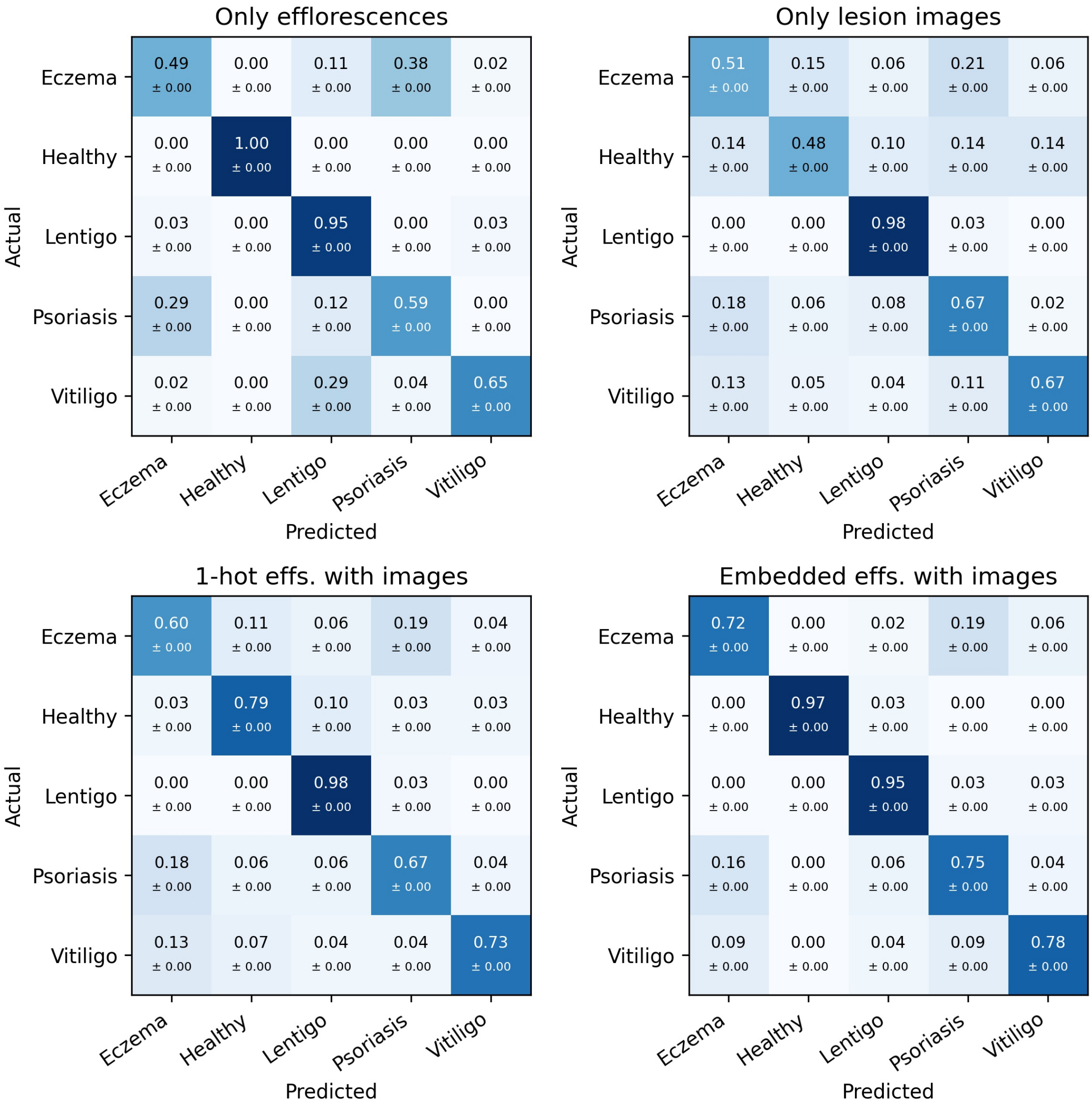


Figure 6.2: Confusion matrices of the differential diagnosis models.

Chapter 7

Severity Grading of Skin Diseases

Severity grading of skin conditions is an important part of dermatology consultations, which determines subsequent decisions including treatment recommendation. It is also the procedure that enables dermatologists to assess the evolution of a disease and the success of their actions (cf. section 2.3). In this chapter, we propose methods to automate the severity evaluation of several diseases based on patients' photographs. Section 7.1 presents our research article on the quantification of palmoplantar pustular psoriasis (PPP) lesions. We show the applicability of our approach in restricted data availability settings by training a deep learning model (DLM) to segment ichthyosis with confetti (IWC) lesions in section 7.2. Finally, we present our research article on hand eczema anatomical stratification in section 7.3.

7.1 Quantification of Efflorescences in Pustular Psoriasis using Deep Learning

This research article was published [10] at the journal of Healthcare Informatics Research [1]. We hypothesized that PPP lesions could be automatically quantified using a segmentation approach at high correlation (> 0.75) with experts' annotations. In this work, we confirm this hypothesis and propose an approach to automatically quantify PPP lesions in terms of counts and surface. The high correlation of our method's predictions with experts' labels shows its potential to improve objectivity and precision of current clinical severity grading systems.

[1] Full text via DOI: `https://doi.org/10.4258/hir.2022.28.3.222` (Accessed: 2nd February 2023)

Original Article

Healthc Inform Res. 2022 July;28(3):222-230.
https://doi.org/10.4258/hir.2022.28.3.222
pISSN 2093-3681 • eISSN 2093-369X

HIR
Healthcare Informatics Research

Quantification of Efflorescences in Pustular Psoriasis Using Deep Learning

Ludovic Amruthalingam[1,4], Oliver Buerzle[2], Philippe Gottfrois[1], Alvaro Gonzalez Jimenez[1], Anastasia Roth[3], Thomas Koller[4], Marc Pouly[4], Alexander A. Navarini[1,5]

[1]Department of Biomedical Engineering, University of Basel, Basel, Switzerland
[2]Department of Dermatology, University Hospital Zurich, Zurich, Switzerland
[3]Department of Health Sciences and Technology, Swiss Federal Institute of Technology, Zurich, Switzerland
[4]Lucerne School of Computer Science and Information Technology, Lucerne University of Applied Sciences and Arts, Lucerne, Switzerland
[5]Department of Dermatology, University Hospital of Basel, Basel, Switzerland

Objectives: Pustular psoriasis (PP) is one of the most severe and chronic skin conditions. Its treatment is difficult, and measurements of its severity are highly dependent on clinicians' experience. Pustules and brown spots are the main efflorescences of the disease and directly correlate with its activity. We propose an automated deep learning model (DLM) to quantify lesions in terms of count and surface percentage from patient photographs. **Methods:** In this retrospective study, two dermatologists and a student labeled 151 photographs of PP patients for pustules and brown spots. The DLM was trained and validated with 121 photographs, keeping 30 photographs as a test set to assess the DLM performance on unseen data. We also evaluated our DLM on 213 unstandardized, out-of-distribution photographs of various pustular disorders (referred to as the pustular set), which were ranked from 0 (no disease) to 4 (very severe) by one dermatologist for disease severity. The agreement between the DLM predictions and experts' labels was evaluated with the intraclass correlation coefficient (ICC) for the test set and Spearman correlation (SC) coefficient for the pustular set. **Results:** On the test set, the DLM achieved an ICC of 0.97 (95% confidence interval [CI], 0.97–0.98) for count and 0.93 (95% CI, 0.92–0.94) for surface percentage. On the pustular set, the DLM reached a SC coefficient of 0.66 (95% CI, 0.60–0.74) for count and 0.80 (95% CI, 0.75–0.83) for surface percentage. **Conclusions:** The proposed method quantifies efflorescences from PP photographs reliably and automatically, enabling a precise and objective evaluation of disease activity.

Keywords: Psoriasis, Dermatology, Computer-Assisted Diagnosis, Machine Learning, Deep Learning

Submitted: March 3, 2022
Revised: April 27, 2022
Accepted: June 29, 2022

Corresponding Author
Alexander A. Navarini
Department of Biomedical Engineering, University of Basel, Petersgraben 4, 4031 Basel, Switzerland. Tel: +41-61-2654080, E-mail: alexander.navarini@usb.ch (https://orcid.org/0000-0001-7059-632X)

I. Introduction

Pustular psoriasis (PP) can impair the quality of life by producing innumerable painful pustules (white or yellow vesicles) on weight-bearing areas, or lead to uncontrollable systemic inflammation and malaise. Both localized and generalized forms exist. Palmoplantar PP (PPP) is the most frequent form and produces numerous pustules on an erythematous base in the palmoplantar region. With time, these pustules dry, and their subsequent secondary efflorescences are termed brown spots. Generalized PP affects the whole body; it is rarer than localized forms and more dangerous in cases with systemic complications. There is no established standard treatment, and the available options are still limited [1].

The severity of a skin disease is traditionally evaluated based on its physical impact on patients' health. Several different metrics exist for psoriasis, of which the Psoriasis Area and Severity Index (PASI) is considered the most established [2]. For PP, there is no universally used grading system. Objective grading systems such as the PPPASI [3] are based on the quantity and intensity of important disease features, most prominently the pustules. As these scoring systems were designed for manual assessment, they use an imprecise grading system from "no disease" (0) to "very severe" (4), integrating pustules, erythema, and scaling. Similarly, the area covered by efflorescences is also graded using discrete categories. Even though such scales are clinically useful and efficient in practice, they clearly constrain precision for severity grading and disease monitoring. As shown by the PrecisePASI for plaque-type psoriasis [4], this limitation can be overcome by developing tools for fine-grained assessments. These precise grading systems are especially important for monitoring patients' conditions and determining the required treatments, as PP is a relapsing disease with varying degrees of severity across flare episodes. Dermatologists usually evaluate PP activity by coarse estimations, which have inevitable disadvantages such as inter-individual variation among raters [5]. Hence, an automated and reliable alternative would benefit clinical practitioners, facilitate medical studies, and could be smoothly integrated into tele-dermatology applications.

In comparison to other inflammatory skin diseases, PP presents distinct and easily identifiable skin lesions: pustules and brown spots. This special characteristic could enable machine learning (ML) algorithms to automatically perform counting and surface estimation, a very daunting task in manual settings. For example, the reader may visually assess the quantity of lesions in the patient's hand shown in Figure 1, which tallies 118 pustules and 272 brown spots and surface percentages 2.11% and 3.14%, respectively. Clearly, such fine-grained assessments can only be achieved through automation.

Current state-of-the-art image recognition models are based on deep learning (DL) architectures. DL is a branch of ML aiming to develop models that autonomously learn relevant discriminating features from data sources to infer predictions on new unseen data samples. These deep learning models (DLMs) can be used in automated pipelines and have the advantage of producing deterministic and therefore reproducible results. They have repeatedly achieved superhuman performance in image recognition tasks, progressing to general images today. Successful applications to medical image analysis include skin cancer classification [6], psoriasis or brain tumor segmentation [7,8] and even synthetic medical data generation [9].

In this study, we propose a DLM to automatically quantify PP efflorescences (lesion count and surface percentage) and evaluate its predictions against experts' labels.

II. Methods

1. PPP Dataset

The dataset consisted of 151 anonymized high-resolution photographs obtained at the University Hospital Zurich from PPP patients with active lesions. Two board-certified dermatologists and a student independently labeled the images for pustules and brown spots. Figure 1 shows an example of a PPP image from our dataset along with its expert labels.

We randomly divided the dataset into 121 photographs to train the DLM and 30 photographs to test its performance, ensuring that the training and test set did not contain any data from the same patient. The training set was further

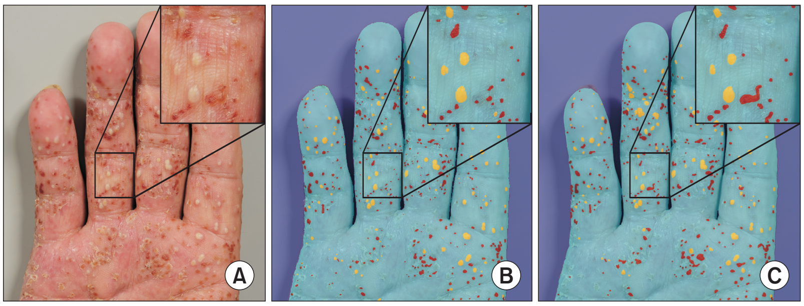


Figure 1. Sample image (A) with expert labels (B) and the DLM prediction (C). This picture came from the test set used to evaluate the DLM and was not used in the training process. The original image is shown in (A), while (B) shows the image overlaid with expert labels and (C) the image overlaid with the DLM predictions. The pustules are colored in yellow, the brown spots in red, the patient's skin in blue, and the background in violet. DLM: deep learning model.

divided into five folds for cross-validation to determine the optimal DLM (hyper-)parameters and to evaluate the variability of the DLM performance across the different training splits.

To leverage the full resolution of the photographs, we tiled the images in square patches with a fixed side length of 512 pixels (approximately 3 cm × 3 cm). This pre-processing step resulted in 6,799 patches for the training set and 819 for the test set. Finally, only the training set was further augmented to improve DLM generalization using random transformations such as flips, rotations, zoom, and contrast and brightness changes. The full test set lesion distribution is displayed in the supplementary materials.

2. DLM Training

The suggested DLM is composed of two subunits, both based on the U-Net [10] architecture with a ResNet [11] backbone to extract image features. The workflow is as follows: first, the M1 subunit separates the skin and background from the full picture, while the M2 subunit splits the picture into patches and segments pustules and brown spots. The M1 predictions take priority over M2 predictions in the sense that we consider M2-predicted pustules and spots only when they overlap with M1-predicted skin. The lesions are counted and the surface percentage (the total lesions' pixel size multiplied by 100, then divided by the total skin's pixel size) is calculated.

Due to the relatively small size of our dataset, the training process was preceded by two pretraining steps. First, we applied transfer learning on both subunits' backbones using the pretrained weights from the ImageNet dataset [12]. Next, we pretrained the M2 subunit's backbone on a simpler classification task: separating patches containing lesions from patches with only background or healthy skin.

Finally the training of the DLM was performed for each subunit independently on the same training set using a learning rate scheduler with a one-cycle policy [13].

As the lesions are very small, there is a large imbalance between lesion pixels and irrelevant pixels from the skin or background. To ensure that the DLM properly learns to recognize very small lesions, we used the mixed focal loss function [14], combining the focal loss [15] and the dice focal loss [16], both of which are known to mitigate semantic class imbalance and are popular in medical image segmentation [17]. The implementation was done with PyTorch [18] and the fastai library [19].

3. Pustular Diseases Dataset (PDD)

This dataset used for out-of-distribution testing consisted of 213 unstandardized pictures from four pustular diseases (Table 1) with at least 15 images per diagnosis (Supplementary Tables S1–S3). The diseases were selected because they also produce pustules and brown spots. One of the four diseases was again PPP, but the pictures were derived from a distinct patient population and were less standardized. In comparison to the training dataset, the PDD pictures varied greatly in terms of resolution, zoom level, focus, brightness level, patient posture, and so on. One dermatologist assessed the images for actual disease severity using a physician's global assessment ranking from 0 (no disease) to 4 (very severe). In contrast, one student graded the images for lesion count only, with results ranging from 0 (no lesions) to 4 (very large count) for the estimated lesion count. Consistent estimation of the lesion surface percentage by human raters was tried, but proved to be too difficult and was therefore abandoned.

4. Analysis

To evaluate the agreement between the experts' labels and the DLM predictions, intraclass correlation coefficients (ICCs) with 95% confidence intervals (CIs) were measured. For the PDD experiment, we computed Spearman correlation (SC) coefficients with a 95% CI instead, since ranking labels are ordinal variables. The computed correlation coefficients reflect how well the DLM predictions relate to the experts' labels: <0.4 for weak agreement, 0.4–0.6 for moderate, 0.61–0.8 for strong, and >0.8 for very strong agreement.

Following the recommendations by van Stralen [20] we created Bland-Altman (BA) plots to analyze the agreement. As the data were not normally distributed, the BA limits of agreements were computed with the 2.5th and 97.5th percentiles (to cover 95% of the data samples). We also created a Q3P plot to show the third quartile of (absolute and relative) differences between experts' labels and DLM predictions.

Table 1. Correlation coefficients of DLM predictions

	ICC	
	Surface	Count
Pustules	0.88 (0.87–0.90)	0.96 (0.96–0.97)
Brown spots	0.92 (0.91–0.93)	0.97 (0.97–0.98)
All lesions	0.93 (0.92–0.94)	0.97 (0.97–0.98)

The values in parenthesis correspond to the 95% confidence interval.

Performance of the deep learning model (DLM) surface and count predictions evaluated on 819 image patches from the test set using the intraclass correlation coefficient (ICC). All *p*-values are below 0.05.

Thus, for both the BA and Q3P plots, a positive difference means that the DLM underestimates the efflorescence quantity while a negative difference implies the opposite.

Finally, in order to better understand the DLM's divergence from the experts' labels, we randomly selected 100 patches from the PPP test set and manually analyzed the lesions missed by the DLM and the lesions that it detected but were missed by the experts. A student then analyzed each case individually and determined if the discrepancy reflected a mistake by the DLM or the experts.

III. Results

The results presented in this section were obtained from the PPP test set patches (Supplementary Figures S1–S3).

1. PPP Test Set: Prediction of Pustule and Brown Spot Counts

As shown in Figure 2F, the DLM predictions differed by at most 1 pustule or brown spot in 75% of the patches with up to 6 lesions (corresponding to the third quartile [Q3] of the test set for lesion count). For the remaining patches (i.e., in 18.8% of all cases), the difference increased to 2 le-

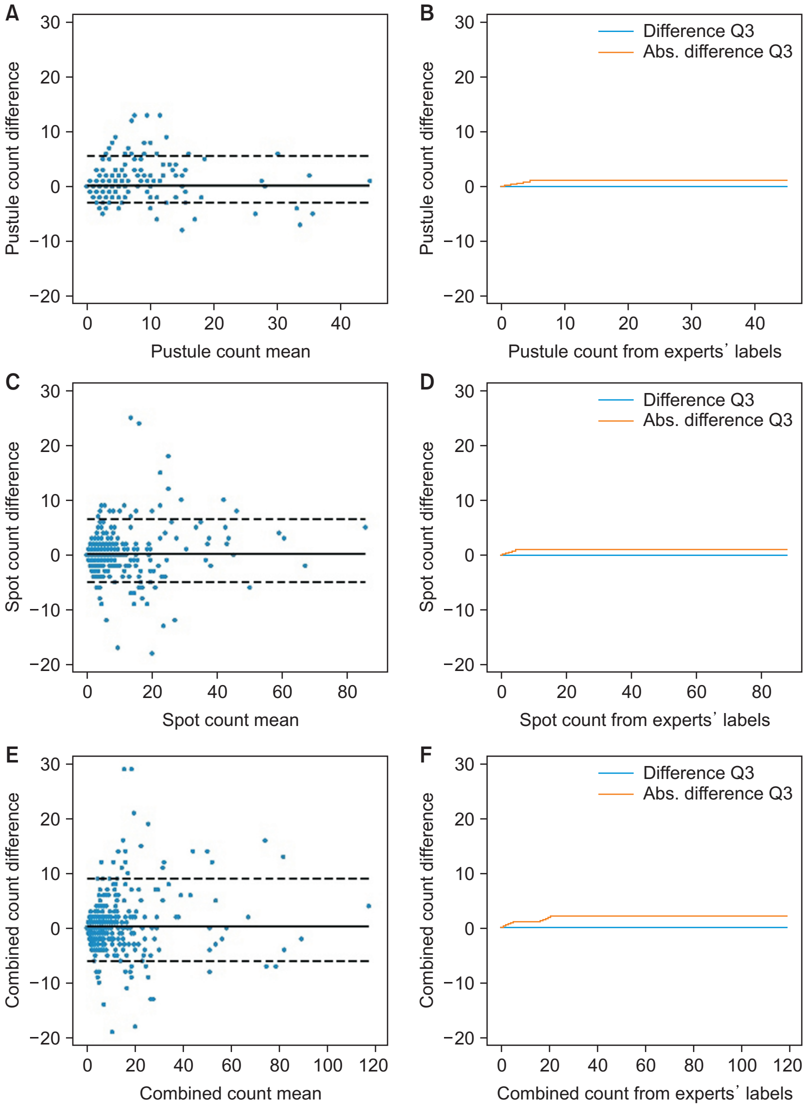


Figure 2. Agreement of DLM lesion count predictions with expert labels. The figure shows the Bland-Altman plots of the predicted count for pustules (A), spots (C), and combined lesions (E). The plots for pustules (B), spots (D), and both lesions (F) show the third quartile of the mean difference and the mean absolute difference of the predicted count for patches with up to the number of lesions specified on the horizontal axis value. DLM: deep learning model.

sions. The DLM's bias (full line in BA plots) was 0.24 lesions (Figure 2E), indicating that the DLM tended to detect fewer lesions than the experts did. The BA plots did not reveal a systematic bias in the DLM predictions; the patches were concentrated on the left of the x-axis because most of them contained only a few lesions. The mean absolute difference (MAD) was 1.68 lesions, and although we observed several outliers, the ICC was 0.97 with (95% CI, 0.97–0.98) (Table 2), implying very strong agreement with the experts' labels.

2. PPP Test Set: Prediction of Pustule and Brown Spot Surface Percentage

Considering the test image patches with lesion surface percentages up to 1.31% (PPP test set's surface Q3), the DLM surface predictions differed by less than 0.15% in 75% of the cases (Figure 3F). This difference plateaued at 0.42% for 75% of the patches with higher surface percentages. The predicted surface ICC was 0.93 with (95% CI, 0.92–0.94) (Table 2). The DLM bias was 0.27% and the MAD was 0.47%, implying that the DLM tended to underestimate the surface of lesions. Again, the BA plots did not reveal any systematic bias in the DLM predictions.

3. PPP Test Set: Review of DLM Divergence

The DLM predictions for all 100 patches yielded 486 lesions, of which 76.6% matched the experts' labels. However, 23.4% were absent from the experts' labels. Manual verification determined that 88.5% were indeed real pustules or brown spots missed by the experts, and only 11.5% were structures mistakenly identified by the DLM.

The experts labeled a total of 579 lesions, of which 63.6% were identified by the DLM, 30.6% were missed, and the remaining 5.8% were upon manual verification identified to be expert label errors; thus, they were correctly classified to be healthy skin by the DLM.

We infer from these observations that from these 100 patches, the correct lesion count should have been 645, implying a combined sensitivity for experts of 84.4% with a labeling error rate of 5.8%, and for the DLM a sensitivity of 73.3% with a detection error rate of 2.6%.

The usual mistakes both for the experts and DLM were caused by lesion-mimicking structures, such as small lentigines or dirt for brown spots and scales for pustules. Concerning the missing lesions from the experts' labels, these were mainly small pustules or brown spots that a human could barely see without sufficient zooming in.

4. PDD Set: DLM Evaluation for Pustular Diseases

We applied the DLM to 213 unstandardized pictures from four different pustular diseases to predict the lesion count and surface. Table 2 shows the corresponding SC coefficients with the experts' grading. With respect to the dermatologist's severity grading, the overall SC coefficient for all diagnoses was 0.66 (95% CI, 0.60–0.74) for lesion count and 0.80 (95% CI, 0.75–0.83) for lesion surface, indicating strong agreement. Regarding the medical student's estimated lesion count, the observed agreement was strong (SC coefficient = 0.77; 95% CI, 0.72–0.81).

IV. Discussion

This work addressed the task of automatically measuring disease intensity in PPP patient photographs. The presented DLM was able to quantify both pustules and brown spots in patient images, reaching very strong agreement with experts' labels, as shown by an ICC range of 0.97–0.98 for lesion count and an ICC range of 0.92–0.94 for lesion surface percentage. An analysis of a randomly selected subsample of the test set revealed a combined expert sensitivity of 84.4% with an error rate of 5.8%, while the DLM showed a sensitivity of

Table 2. Pustular diseases dataset

Diagnosis	Spearman correlation coefficient		
	Surface A	Count A	Count B
All diagnoses	0.80 (0.75–0.83)	0.66 (0.60–0.74)	0.77 (0.72–0.81)
Acropustulosis of infancy	0.83 (0.61–0.96)	0.71 (0.50–0.92)	0.66 (0.31–0.89)
Palmoplantar pustular psoriasis	0.76 (0.69–0.85)	0.70 (0.60–0.79)	0.78 (0.73–0.86)
Pustulosis palmoplantaris	0.78 (0.70–0.85)	0.67 (0.52–0.79)	0.74 (0.63–0.84)
Pustulosis subcornealis	0.75 (0.60–0.82)	0.75 (0.61–0.87)	0.87 (0.82–0.91)

The values in parenthesis correspond to the 95% confidence interval.

Performance of the deep learning model (DLM) surface and count predictions evaluated on the 213 images from the pustular disease dataset with the Spearman correlation coefficients. The columns labeled A correspond the dermatologist's disease severity ranking and B, the medical student's lesion count ranking. All *p*-values are below 0.05.

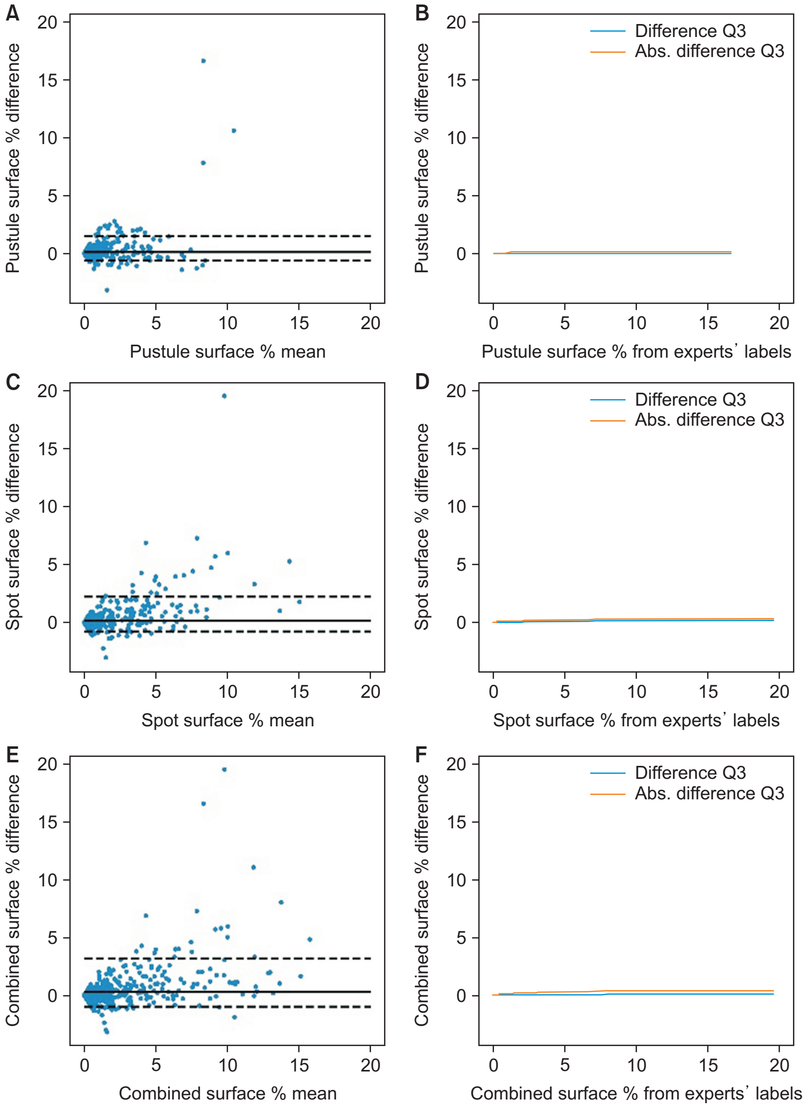


Figure 3. Agreement of DLM lesion surface predictions with expert labels. The figure shows the Bland-Altman plots of the predicted surface percentage for pustules (A), spots (C) and combined lesions (E). The plots for pustules (B), spots (D), and both lesions (F) show the third quartile of the mean difference and the mean absolute difference of the predicted surface percentage for patches with up to the lesion surface specified on the horizontal axis value. DLM: deep learning model.

73.3% with an error rate of 2.6%.

The DLM was further evaluated on photographs taken from patients with four pustular diseases. It showed strong agreement with the dermatologist's severity evaluation (on a range from 0 to 4) and the student's lesion count (likewise on a scale from 0 to 4). To the best of our knowledge, this is the first attempt to automatically quantify efflorescences from pustular psoriasis; as such, this is the first step toward a precise, reproducible, and objective evaluation of this disease activity.

Related to the task of automating existing disease scoring systems, most of the literature has focused on the automation of the PASI index. Some studies [21-23] chose to rely on classification DLMs, thus capping the achievable precision to discrete scores in contrast to our DLM, which predicts continuous metrics. Various segmentation approaches have also been applied to ulcers [24], skin cancer [25,26], eczema [27], and psoriasis [7,28], and therefore could also be used to produce metrics similar to our study. However, they all targeted diseases with plaques, single lesions, or lesions larger than PP efflorescences. The segmentation of small objects in imbalanced settings is a well-known technical challenge [29], which we successfully addressed here in the context of PP with our patch-based approach and an additional pretraining task. This patch-based approach was the main motivation behind our design choice to segment skin separately from lesions, since the first task is performed better when the full image context is available. Another PP-specific difficulty was caused by the inevitably limited sensitivity of experts in cases with a large number of lesions and the tedious nature of the labeling task. To illustrate the impact on the clinical workload, the image shown in Figure 1 required 30 minutes for the human expert to fully label, whilst the same took less than 15 seconds for the DLM. The produced labels were bound to miss some lesions, penalizing the DLM training and evaluation process. Indeed when analyzing the quantitative DLM segmentation performance (see Supplementary Figures S4 and S5), around 40% of lesion pixels were mistaken for healthy skin, matching the observed positive bias in the counts and surface Bland-Altman plots. However, the high intra-class correlation with experts' labels implies that the disease lesions were quantified according to the experts' annotations, aligning with the study's main objective.

Due to its algorithmic nature, the error rate of the DLM should remain constant in time across different patient cases. We expect the DLM's performance to be at least as stable as human evaluation over the course of various follow-up visits. Both hypotheses should be validated in future studies.

While our DLM was trained exclusively on PPP patients' pictures, we demonstrated that our approach of counting lesions and measuring their surface to evaluate the disease severity is also applicable to relatively unstandardized, out-of-distribution (coming from a different source with different capturing conditions) photographs of patients with other pustular disorders.

This remarkable generalization is possible without retraining the DLM as long as the different diseases' lesions have a similar appearance. Whilst the pictures showed very different patient postures and body regions, the DLM's performance remained robust, presumably due to its training on small image patches instead of full images.

Dermatologists' workflow currently consists of either an informal subjective global assessment or manually grading disease activity with an objective score such as the PPPASI. The latter, however, requires time and expertise to perform in a reproducible manner. Improving on this situation, our approach for PP grading does not have such constraints. The DLM could be integrated into a smartphone app enabling physician extenders to photograph and quantify lesions before patients consult with dermatologists. To allow a systematic comparison of the DLM predictions, it is important to standardize the conditions under which pictures are taken, such as a patient's posture, zoom level, and so forth. This could be achieved via a guided picture-taking process in the smartphone app and proper training of medical personnel.

Image standardization is a common pitfall for DLMs. When photographs are taken with very different settings (lighting, posture, or zoom level), the quality of DLM predictions can degrade despite training with extensive data augmentation. Such variations can be reduced by following photograph collection procedures such as the guidelines proposed by Finnane et al. [30] for dermatology. Although our DLM showed robust performance on unstandardized pictures, they were taken by photographers and medical personnel in relatively controlled conditions (hospitals and studies). For extreme cases such as tele-dermatology (where untrained people take images with different devices, resolution, zoom, exposure to sunlight, and so forth) the DLM should be retrained using transfer learning on a subset of the new data source. Another limitation to consider is that the DLM was trained in this study mainly with Caucasian patient pictures and must therefore be retrained before it is applied to patients with different skin pigmentation. Once a new dataset has been collected, DLM retraining is usually not a challenging task since it is possible to leverage the already learned knowledge with transfer learning.

Another common criticism of DL applications in medicine is the difficulty of explaining the rationale behind model predictions, which makes them unsafe for use in tasks such as differential diagnosis. Here, this issue is not critical since the presented approach can be validated with little effort and training by visualizing the predicted lesions (a single glance would be sufficient).

Our DLM enables new, previously impractical analyses, including systematic studies of pustules' growth, shapes, evolution, and treatment response. In practice, our approach is particularly suited for automatically generating patient reports, disease monitoring, and analyzing treatment efficacy. It synergizes well with standardized full-body photography solutions and their respective image analysis pipelines. In the future, our method could be utilized to develop tools that would help dermatologists better monitor patients afflicted with any type of pustulosis or disseminated monomorphic rashes and therefore improve the quality of follow-up consultations. The DLM is well-suited for integration into tele-dermatology applications, provided it is retrained to match the expected types of inputs and complemented with systems to ensure picture quality and verify the output. This could reduce hospital loads and be deployed in geographical regions where physical access to dermatologists is difficult or even impossible.

Conflict of Interest

Alexander 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, Bristol Myers Squibb, Boehringer Ingelheim, Celgene, Eli Lilly, Galderma, GlaxoSmithKline, LEO Pharma, Janssen-Cilag, MSD, Novartis, Pfizer, Pierre Fabre Pharma, Regeneron, Sandoz, Sanofi, and UCB. None of the activities listed above had an impact on this work. All other authors declare no potential conflicts of interest.

Acknowledgments

We thank the members of the labelling consortium: Dr. Komal Agarwal, Dr. Joanna Goldberg, Dr. Swathi Shivakumar, Nicholas Khoury, Anke Naedele. This work was supported by the Helmut-Fischer Foundation, the Botnar Foundation and the University of Basel.

ORCID

Ludovic Amruthalingam (https://orcid.org/0000-0001-5980-5469)
Oliver Buerzle (https://orcid.org/0000-0002-3036-8450)
Philippe Gottfrois (https://orcid.org/0000-0001-8023-3207)
Alvaro Gonzalez Jimenez (https://orcid.org/0000-0002-1337-9430)
Anastasia Roth (https://orcid.org/0000-0003-3199-1006)
Thomas Koller (https://orcid.org/0000-0003-2309-5359)
Marc Pouly (https://orcid.org/0000-0002-9520-4799)
Alexander A. Navarini (https://orcid.org/0000-0001-7059-632X)

Supplementary Materials

Supplementary materials can be found via https://doi.org/10.4258/hir.2022.28.3.222.

References

1. Gooderham MJ, Van Voorhees AS, Lebwohl MG. An update on generalized pustular psoriasis. Expert Rev Clin Immunol 2019;15(9):907-19. https://doi.org/10.1080/1744666X.2019.1648209
2. Puzenat E, Bronsard V, Prey S, Gourraud PA, Aractingi S, Bagot M, et al. What are the best outcome measures for assessing plaque psoriasis severity? A systematic review of the literature. J Eur Acad Dermatol Venereol 2010;24 Suppl 2:10-6. https://doi.org/10.1111/j.1468-3083.2009.03562.x
3. Bhushan M, Burden AD, McElhone K, James R, Vanhoutte FP, Griffiths CE. Oral liarozole in the treatment of palmoplantar pustular psoriasis: a randomized, double-blind, placebo-controlled study. Br J Dermatol 2001;145(4):546-53. https://doi.org/10.1046/j.1365-2133.2001.04411.x
4. Kolios AG, French LE, Navarini AA. Detection of small changes in psoriasis intensity with PrecisePASI. Dermatology 2015;230(4):314-7. https://doi.org/10.1159/000371811
5. Youn SW, Choi CW, Kim BR, Chae JB. Reduction of inter-rater and intra-rater variability in psoriasis area and severity index assessment by photographic training. Ann Dermatol 2015;27(5):557-62. https://doi.org/10.5021/ad.2015.27.5.557
6. Esteva A, Kuprel B, Novoa RA, Ko J, Swetter SM, Blau HM, et al. Dermatologist-level classification of skin cancer with deep neural networks. Nature 2017;542(7639):115-8. https://doi.org/10.1038/nature21056
7. Meienberger N, Anzengruber F, Amruthalingam L, Christen R, Koller T, Maul JT, et al. Observer-independent assessment of psoriasis-affected area using machine learning. J Eur Acad Dermatol Venereol 2020;34(6):1362-8. https://doi.org/10.1111/jdv.16002
8. Andermatt S, Horvath A, Pezold S, Cattin P. Pathology segmentation using distributional differences to images of healthy origin. In: Crimi A, Bakas S, Kuijf H, Keyvan F, Reyes M, van Walsum T, editors. Brainlesion: glioma, multiple sclerosis, stroke and traumatic brain injuries. Cham, Switzerland: Springer; 2018. p. 228-38. https://doi.org/10.1007/978-3-030-11723-8_23
9. Furger F, Amruthalingam L, Navarini A, Pouly M. Applications of generative adversarial networks to dermatologic imaging. In: Schilling FP, Stadelmann T, editors. Artificial neural networks in pattern recognition. Cham, Switzerland: Springer; 2020. p. 187-99. https://doi.org/10.1007/978-3-030-58309-5_15
10. Ronneberger O, Fischer P, Brox T. U-Net: convolutional networks for biomedical image segmentation. In: Navab N, Hornegger J, Wells W, Frangi A, editors. Medical image computing and computer-assisted intervention – MICCAI 2015. Cham, Switzerland: Springer; 2015. p. 234-41. https://doi.org/10.1007/978-3-319-24574-4_28
11. He K, Zhang X, Ren S, Sun J. Deep residual learning for image recognition. In Proceedings of the IEEE conference on computer vision and pattern recognition; 2016 Jun 26-Jul 1; Las Vegas, NV. p. 770-8.

12. Deng J, Dong W, Socher R, Li LJ, Li K, Fei-Fei L. ImageNet: a large-scale hierarchical image database. Proceedings of 2009 IEEE Conference on Computer Vision and Pattern Recognition; 2009 Jun 20-25; Miami, FL. p. 248-55. https://doi.org/10.1109/CVPR.2009.5206848
13. Smith LN. A disciplined approach to neural network hyper-parameters: Part 1--learning rate, batch size, momentum, and weight decay [Internet]. Ithaca (NY): arXiv.org; 2018 [cited at 2022 Jul 20]. Available from: https://arxiv.org/abs/1803.09820.
14. Yeung M, Sala E, Schonlieb CB, Rundo L. Unified focal loss: generalising dice and cross entropy-based losses to handle class imbalanced medical image segmentation. Comput Med Imaging Graph 2022;95:102026. https://doi.org/10.1016/j.compmedimag.2021.102026
15. Lin TY, Goyal P, Girshick R, He K, Dollar P. Focal loss for dense object detection. IEEE Trans Pattern Anal Mach Intell 2020;42(2):318-27. https://doi.org/10.1109/TPAMI.2018.2858826
16. Zhu W, Huang Y, Zeng L, Chen X, Liu Y, Qian Z, et al. AnatomyNet: deep learning for fast and fully automated whole-volume segmentation of head and neck anatomy. Med Phys 2019;46(2):576-89. https://doi.org/10.1002/mp.13300
17. El Jurdi R, Petitjean C, Honeine P, Cheplygina V, Abdallah F. High-level prior-based loss functions for medical image segmentation: a survey. Comput Vis Image Underst 2021;210:103248. https://doi.org/10.1016/j.cviu.2021.103248
18. Paszke A, Gross S, Massa F, Lerer A, Bradbury J, Chanan G, et al. PyTorch: an imperative style, high-performance deep learning library. Adv Neural Inf Process Syst 2019;32:8024-35.
19. Howard J, Gugger S. Fastai: a layered API for deep learning. Information 2020;11(2):108. https://doi.org/10.3390/info11020108
20. van Stralen KJ, Dekker FW, Zoccali C, Jager KJ. Measuring agreement, more complicated than it seems. Nephron Clin Pract 2012;120(3):c162-7. https://doi.org/10.1159/000337798
21. Schaap MJ, Cardozo NJ, Patel A, de Jong EM, van Ginneken B, Seyger MM. Image-based automated psoriasis area severity index scoring by convolutional neural networks. J Eur Acad Dermatol Venereol 2022;36(1):68-75. https://doi.org/10.1111/jdv.17711
22. Wu X, Yan Y, Zhao S, Kuang Y, Ge S, Wang K, et al. Automatic severity rating for improved psoriasis treatment. In: Medical image computing and computer assisted intervention – MICCAI 2021. Cham, Switzerland: Springer; 2021. p. 185-94. https://doi.org/10.1007/978-3-030-87234-2_18
23. Pal A, Chaturvedi A, Garain U, Chandra A, Chatterjee R, Senapati S. Severity assessment of psoriatic plaques using deep CNN based ordinal classification. In: OR 2.0 Context-aware operating theaters, computer assisted robotic endoscopy, clinical image-based procedures, and skin image analysis. Cham, Switzerland: Springer; 2018. p. 252-9. https://doi.org/10.1007/978-3-030-01201-4_27
24. Cazzolato MT, Ramos JS, Rodrigues LS, Scabora LC, Chino DY, Jorge AE, et al. The UTrack framework for segmenting and measuring dermatological ulcers through telemedicine. Comput Biol Med 2021;134:104489. https://doi.org/10.1016/j.compbiomed.2021.104489
25. Zhao C, Shuai R, Ma L, Liu W, Wu M. Segmentation of dermoscopy images based on deformable 3D convolution and ResU-NeXt +. Med Biol Eng Comput 2021;59(9):1815-32. https://doi.org/10.1007/s11517-021-02397-9
26. Goyal M, Oakley A, Bansal P, Dancey D, Yap MH. Skin lesion segmentation in dermoscopic images with ensemble deep learning methods. IEEE Access 2019;8:4171-81. https://doi.org/10.1109/ACCESS.2019.2960504
27. Schnurle S, Pouly M, vor der Bruck T, Navarini A, Koller T. On using support vector machines for the detection and quantification of hand eczema. Proceedings of the 9th International Conference on Agents and Artificial Intelligence (ICAART); 2017 Feb 24-26; Porto, Portugal. p. 75-84.
28. Raj R, Londhe ND, Sonawane RS. Deep learning based multi-segmentation for automatic estimation of psoriasis area score. Proceedings of 2021 8th International Conference on Signal Processing and Integrated Networks (SPIN); 2021 Aug 26-27; Noida, India. p. 1137-42. https://doi.org/10.1109/SPIN52536.2021.9566039
29. Liu Y, Sun P, Wergeles N, Shang Y. A survey and performance evaluation of deep learning methods for small object detection. Expert Syst Appl 2021;172:114602. https://doi.org/10.1016/j.eswa.2021.114602
30. Finnane A, Curiel-Lewandrowski C, Wimberley G, Caffery L, Katragadda C, Halpern A, et al. Proposed technical guidelines for the acquisition of clinical images of skin-related conditions. JAMA Dermatol 2017;153(5):453-7. https://doi.org/10.1001/jamadermatol.2016.6214

Supplementary Materials

Healthc Inform Res. 2022 July;28(3):00-00.
https://doi.org/10.4258/hir.2022.28.3.1
pISSN 2093-3681 • eISSN 2093-369X

HIR
Healthcare Informatics Research

Table S1. Dermatologists' disease severity grading (A) distribution

Diagnosis	S0	S1	S2	S3	S4	Total
Acropustulosis of infancy	1	9	5	2	0	17
Palmoplantar pustular psoriasis	11	20	21	39	4	95
Pustulosis palmoplantaris	0	27	20	14	0	61
Pustulosis subcornealis	0	9	19	12	0	40
All diagnoses	12	65	65	67	4	213

Table S2. Medical student's lesion count ranking (B) distribution

Diagnosis	S0	S1	S2	S3	S4	Total
Acropustulosis of infancy	2	7	6	1	1	17
Palmoplantar pustular psoriasis	10	35	21	21	8	95
Pustulosis palmoplantaris	0	27	21	8	5	61
Pustulosis subcornealis	0	20	6	10	4	40
All diagnoses	12	89	54	40	18	213

Table S3. Correlation coefficients of predictions aggregated on full images

	ICC	
	Surface	Count
Pustules	0.94 (0.89–0.97)	0.99 (0.98–1.00)
Brown spots	0.96 (0.93–0.99)	0.98 (0.98–0.99)
All lesions	0.98 (0.96–0.99)	0.99 (0.98–1.00)

The values in parentheses correspond to the 95% confidence interval. Results obtained from the 30 images in the test set.

ICC: intraclass correlation coefficient.

All *p*-values are below 0.05.

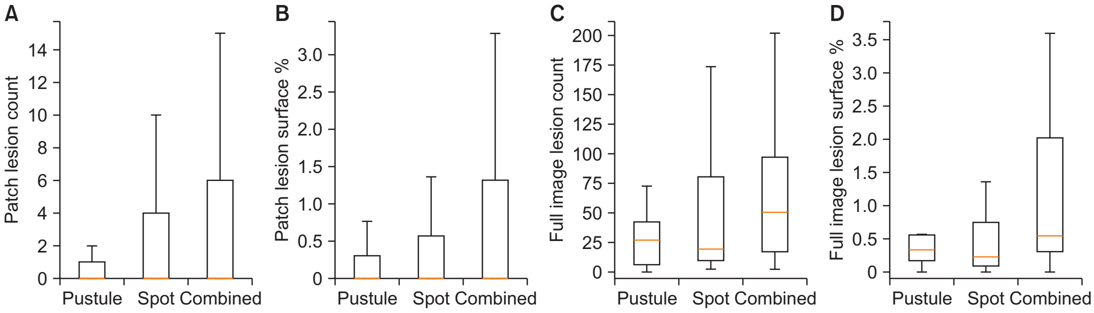


Figure S1. PPP test set lesion distribution. Plots (A) and (B) show, respectively, the count and surface distribution for image patches in the test set. Plots (C) and (D) show the same for the corresponding full images. PPP, palmoplantar pustular psoriasis.

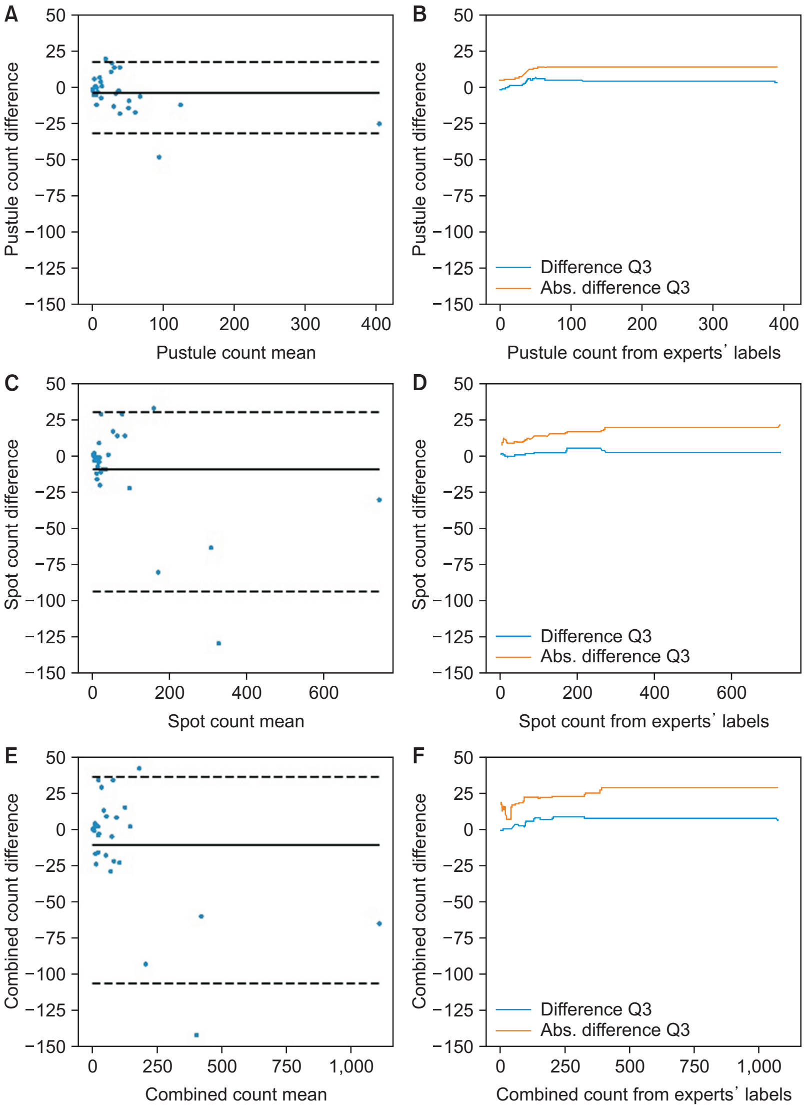


Figure S2. Agreement of count predictions with expert labels on full images. The DLM predictions differed by at most 22.5 lesions in 75% of the patches with up to 97 lesions (the test set's Q3). For the remaining patches, the difference increased to 29 lesions in 75% of the cases. The DLM's bias was -11.1 for both types of lesions, its MAD was 23.96, and the ICC was 0.99 (95% CI, 0.98-1.00). DLM: deep learning model, MAD: mean absolute difference, ICC: intraclass correlation coefficient, CI: confidence interval.

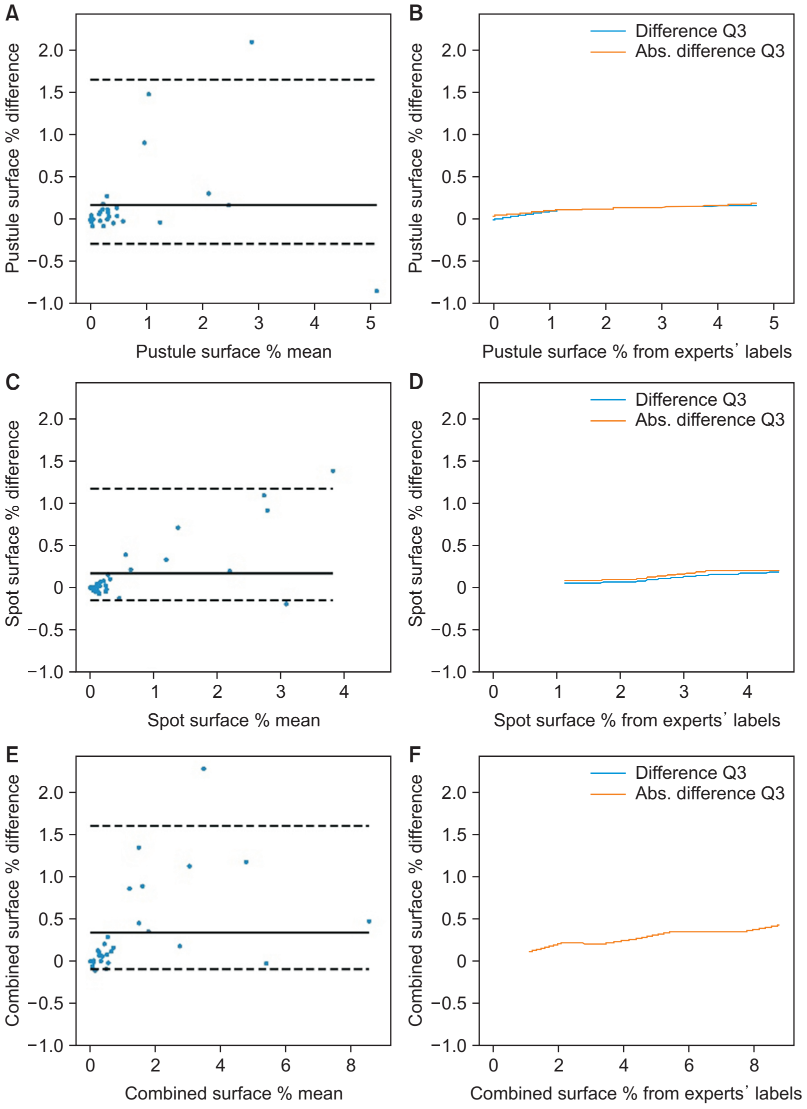


Figure S3. Agreement of surface predictions with expert labels on full images. Considering the test image patches with up to 2% (the test set Q3) of the skin surface covered by pustules and brown spots, the DLM was able to determine the surface with less than 0.22% difference from dermatologists in 75% of the cases. This difference plateaued at 0.42% for 75% of the images with higher surface percentages. The predicted surface ratios of lesions related to the experts' labels with an ICC of 0.98 (95% CI, 0.96–0.99). The DLM bias was 0.33% while the MAD was 0.35%. DLM: deep learning model, MAD: mean absolute difference, ICC: intraclass correlation coefficient, CI: confidence interval.

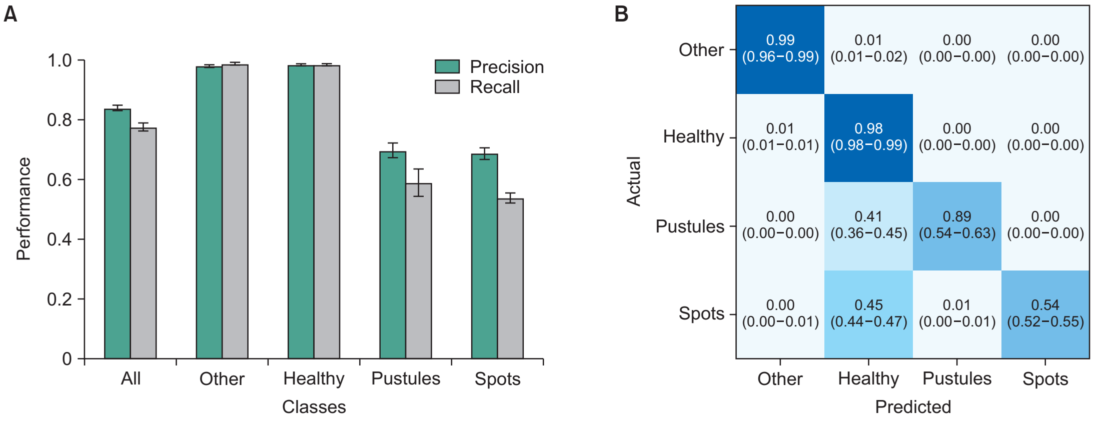


Figure S4. Pixel-wise performance of the DLM in segmentation. Plot (A) shows the pixel precision and recall reached on the test set by the DLM. The first two bars, for the “all” category, represent the macro average of the classes’ individual performance. Plot (B) is a confusion matrix showing the mean proportion of pixels classified among the different classes. Its vertical axis represents the true pixel labels, while the horizontal axis shows the predicted labels. The error bars and values in parentheses represent the 95% confidence interval. The evaluation of the DLM’s pixel-wise performance showed a precision and a recall of 69% and 59% respectively for pustules, and 68% and 54% for brown spots. The DLM missed 41% of pustules pixels and 45% of brown spots pixels, matching the previous observation that it underestimated the lesion sizes. These relatively low scores are a direct consequence of the idiosyncrasy of the experts’ labels. We also evaluated the segmentation performance without ImageNet pretraining and observed a drop in performance. For pustules, we calculated a precision of 35% and recall of 36%, while for brown spots the precision was 48% and the recall was 47%. According to the DLM hyperparameters, with cross-validation, we selected the following hyperparameters for both skin and lesion segmentation DLMs: the batch size was 16, the initial learning rate was 1e-4, the input size was 380 × 380 pixels, and the number of epochs was 40. DLM: deep learning model.

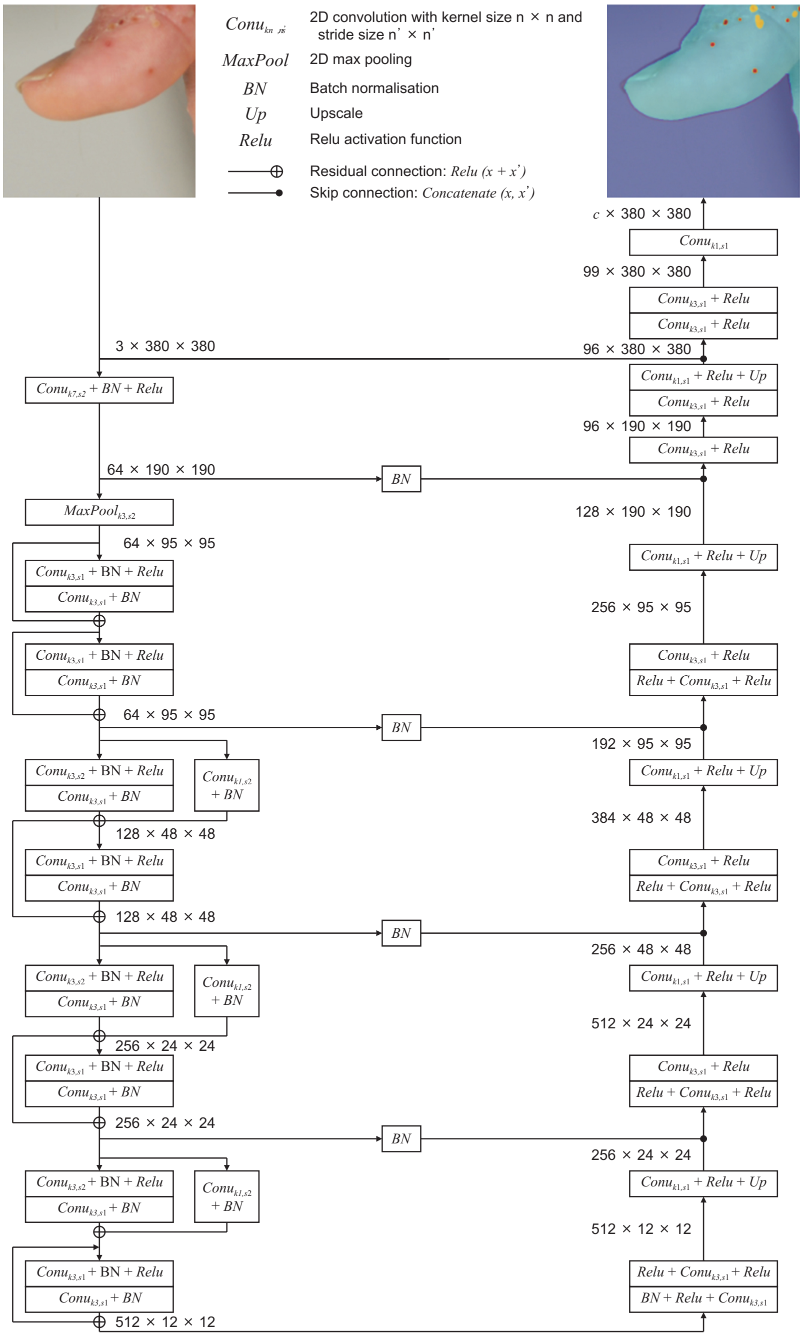


Figure S5. Architecture of the deep learning segmentation model. This figure presents the structure of the segmentation models, based on the U-Net and ResNet architectures. The final mask channel dimension was c = 2 for skin segmentation (M1 subunit) and c = 3 for lesion segmentation (M2 subunit).

7.2 Segmentation of Ichthyosis with Confetti Lesions

IWC is a congenital disease causing the skin to thicken and develop a red appearance as well as different clinical signs including ear deformities or excessive hair growth. The disease is persistent, although its severity evolves throughout patients' growth [23]. It is recognizable by its white spots, which correspond to regions of the skin that spontaneously healed from the ichthyosis. This healing is limited but durable and white spot's appearance remains stable excepting external transformation such as weight gain. Only 40 cases were reported worldwide [69] making IWC a very rare disease with little available data. Currently, no treatment exists, and the disease is an open topic of research. Researchers aim to better understand the disease and follow its evolution. The quantity and anatomical distribution of white spots are useful information in this regard. However, they are too numerous to be precisely estimated manually.

In this section, we present an approach (inspired from section 7.1) to automatically segment white skin and red skin in IWC patients pictures. These results were not published yet.

7.2.1 Materials and Methods

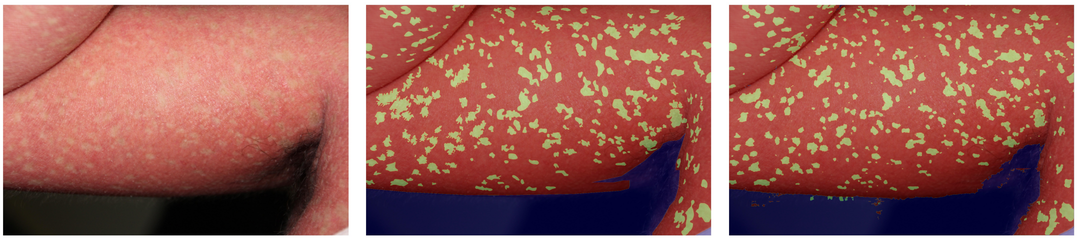
Due to the rarity of the disease, only thirty-five high-resolution pictures of IWC patients were available. Pictures were provided by Dr. Bettina Burger and her team. The images were annotated by a medical student for background, red skin and white skin and then divided into square patches of side-size 512 pixels. Setting aside six images for testing, we trained a ResNet18 [77] following the approach described in section 7.1 for PPP lesions.

7.2.2 Results

The DLM achieved a precision of 81% (95% confidence interval 73-85) and a sensitivity of 80% (CI 73-84) (table 7.1). One test sample with student's annotations and DLM predictions is shown in figure 7.1. Furthermore, we also present predictions for out-of-distribution pictures scanned from research articles in figure 7.2.

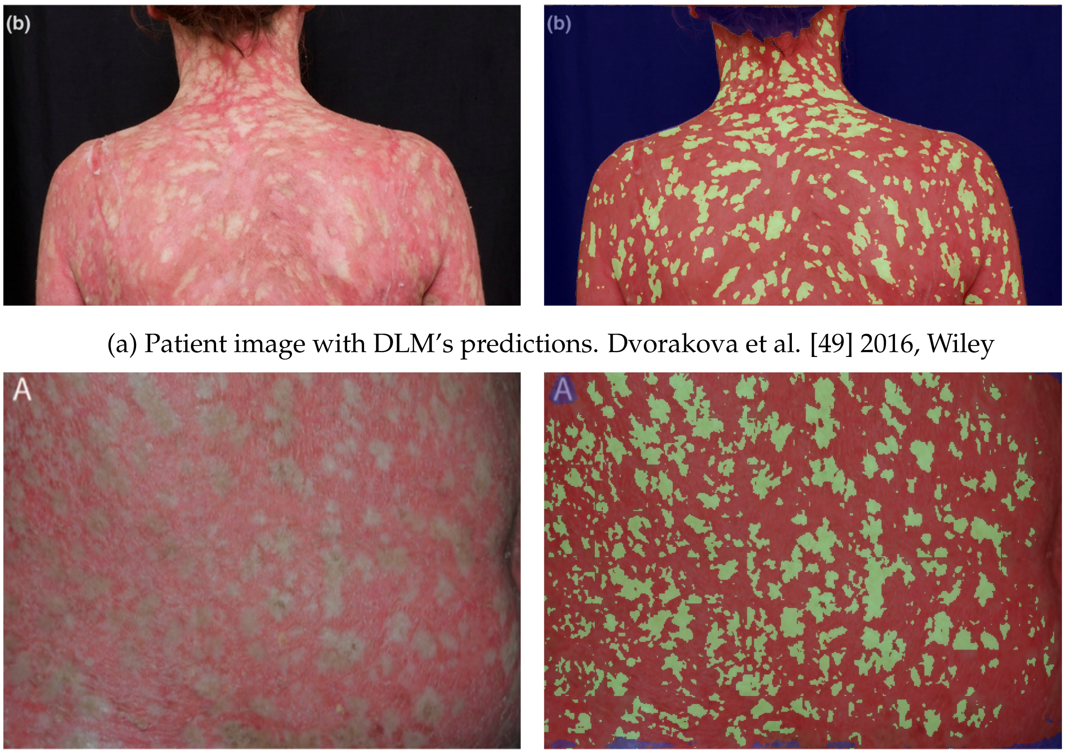
Region	**Precision**	**Sensitivity**
Background	98% (96-99)	94% (90-97)
White spot	81% (73-85)	80% (73-84)
Red skin	93% (89-97)	95% (92-97)
Average	91% (89-92)	90% (88-90)

Table 7.1: Performance of the IWC segmentation DLM.



(a) Patient's image (b) Expert's labels (c) DLM's predictions

Figure 7.1: Test sample of the IWC dataset.



(a) Patient image with DLM's predictions. Dvorakova et al. [49] 2016, Wiley

(b) Patient image with DLM's predictions. Hotz et al. [83] 2016, Medical Journals Sweden AB.

Figure 7.2: Predictions of out-of-distribution samples from previous publications.

7.2.3 Discussion

The high performance of the DLM suggests that the proposed training approach is applicable also in restricted data availability regimes. This is confirmed by the quality of predictions on out-of-distribution samples, which illustrates the generalization capacity of the trained DLM. The performance's large confidence intervals can be explained by the small size of the test dataset. Our approach enables researcher to automatically determine the surface and counts of white spots in patients' photographs. Combined with the anatomy DLMs of chapter 5, their anatomical distribution could also be evaluated similarly to the anatomical stratification of hand eczema in section 7.3.

7.3 Objective Hand Eczema Severity Assessment with Automated Lesion Anatomical Stratification

This research article was published [12] at the journal of Experimental Dermatology [2]. Our work was based on two hypotheses. First, we hypothesized that the surface of hand eczema lesions could be automatically estimated using a segmentation approach at high correlation (> 0.75) with experts' annotations. Second, we hypothesized that the anatomical distribution of hand eczema lesions could be automatically determined.

Both hypotheses were confirmed: we combine our anatomy mapping approach (cf. section 5.2) with our lesion segmentation method (cf. section 7.1) to automatically stratify the surface of hand eczema lesions on the anatomical subregions of the hand. This approach enables objective and precise assessment of the disease and illustrates how features from dermatological lesion description can synergize with deep learning applications in dermatology.

[2] Full text via DOI: `https://doi.org/10.1111/exd.14744` (Accessed: 2nd February 2023)

Received: 5 June 2022 | Revised: 23 December 2022 | Accepted: 7 January 2023

DOI: 10.1111/exd.14744

RESEARCH ARTICLE

Experimental Dermatology WILEY

Objective hand eczema severity assessment with automated lesion anatomical stratification

Ludovic Amruthalingam[1,2] | Nora Mang[3] | Philippe Gottfrois[1] | Alvaro Gonzalez Jimenez[1] | Julia-Tatjana Maul[4,5] | Michael Kunz[3] | Marc Pouly[2] | Alexander A. Navarini[1,3]

[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 Dermatology, University Hospital of Basel, Basel, Switzerland

[4]Department of Dermatology, University Hospital Zurich, Zurich, Switzerland

[5]Faculty of Medicine, University of Zürich, Zurich, Switzerland

Correspondence

Alexander A. Navarini, Department of Dermatology, University Hospital of Basel, Basel, Switzerland

Email: alexander.navarini@usb.ch

Funding information

Fondation Botnar; Helmut Fischer Stiftung; Universität Basel

Abstract

Hand eczema (HE) is one of the most frequent dermatoses, known to be both relapsing and remitting. Regular and precise evaluation of the disease severity is key for treatment management. Current scoring systems such as the hand eczema severity index (HECSI) suffer from intra- and inter-observer variance. We propose an automated system based on deep learning models (DLM) to quantify HE lesions' surface and determine their anatomical stratification. In this retrospective study, a team of 11 experienced dermatologists annotated eczema lesions in 312 HE pictures, and a medical student created anatomical maps of 215 hands pictures based on 37 anatomical subregions. Each data set was split into training and test pictures and used to train and evaluate two DLMs, one for anatomical mapping, the other for HE lesions segmentation. On the respective test sets, the anatomy DLM achieved average precision and sensitivity of 83% (95% confidence interval [CI] 80–85) and 85% (CI 82–88), while the HE DLM achieved precision and sensitivity of 75% (CI 64–82) and 69% (CI 55–81). The intraclass correlation of the predicted HE surface with dermatologists' estimated surface was 0.94 (CI 0.90–0.96). The proposed method automatically predicts the anatomical stratification of HE lesions' surface and can serve as support to evaluate hand eczema severity, improving reliability, precision and efficiency over manual assessment. Furthermore, the anatomical DLM is not limited to HE and can be applied to any other skin disease occurring on the hands such as lentigo or psoriasis.

KEYWORDS

anatomy, computer-assisted, diagnosis, deep learning, eczema, severity of illness index

1 | INTRODUCTION

Hand eczema (HE), also called hand dermatitis, is an inflammatory disease, often chronic, causing a wide spectrum of symptoms including redness (erythema), scaling, hyperkeratosis, fissures, vesicles and erosions.[1] All these features are visible on digital pictures. It is one of the most frequent dermatoses with 15% life prevalence and 10% 1-year prevalence in the general population. It has a multifactorial aetiology including both environmental and genetic factors.[2] HE severity range spans from mild to severe cases, the latter causing adverse physical and psychological effects, both in private and professional activities, and significant impairment to patients' quality of life.[3,4] Globally, HE induced socio-economic burden on society is considerable.[5]

Hand eczema is a remitting and relapsing disease that can acutely flare but also persist in a chronic form. The majority of cases are characterized as occupational, and although current treatments may improve patient's conditions, the disease remains very often chronic, oscillating between acute and subacute stages.[6,7] It is therefore critical that clinicians can monitor its evolution precisely and efficiently to adapt treatment in consequence. A review reported the use of 45 different grading systems in HE research studies.[8] While all of them were based on a selection of morphological patterns and physiological abnormalities, the most accurate in terms of lesion distribution analysed the subregions of each hand separately. One of the most established systems is the hand eczema severity index (HECSI),[9] which consists in combining the rankings of six clinical signs (erythema, induration/papulation, vesicles, fissures, scaling and oedema) with the estimated surface of eczema lesions on five hand subregions (fingertips, fingers without tips, palm of hands, back of hands and wrist). The large variety of clinical signs is caused by the existence of many subtypes of the condition such as dry fissured HE, pulpitis HE, nummular HE, vesicular HE and hyperkeratotic palmar HE.[1]

In clinical practice, severity grading is not performed systematically as it is a time-consuming process (especially when grading is not performed on a regular basis) requiring both training and experience. An overall acute or chronic, mild, moderate to severe grading, eventually with photo documentation, is preferred instead. In situations where precise assessment is required such as the evaluation for fitness to work or reimbursement for expensive drugs, more objective methods like the HECSI score should be performed. Such assessments and the monitoring of disease evolution can only be performed on patients' follow-up (in-person) visits by trained clinicians. Furthermore, precision remains bounded by the discrete nature of the rankings, which induce inevitable inter- and intra-observer variations.[9] This issue was recently illustrated by two independent studies, which reported remarkably different minimal important change values for the HECSI score (41 points[10] vs. 6.3 points[11] on a theoretical maximum of 360 points).

Machine learning algorithms have the potential to assist clinicians with HE severity assessment and monitoring, improving on the efficiency, precision and simplicity of the process. Being automated, they are reproducible and promise to reduce the problem of inter- and intra-observer variations. The best results for machine vision are currently achieved with deep learning models[12,13] (DLM). In this study, we trained two separate DLMs to automatically segment HE lesions and generate the anatomical maps of patients' hands pictures. By combining these predictions, we could generate the anatomical repartition of HE lesions, which can assist with patient documentation and support the determination of severity gradings such as HECSI score.

2 | METHODS

All hand pictures were obtained at the university hospital of Zurich from adult patients, skin type 1 to type 3 on the Fitzpatrick scale over a period of 4 years starting in 2014. The hospital's dermatologists diagnosed patients with HE lesions and then sent them for imaging. Pictures were captured within the same hospital using either a dedicated device under nurse supervision (a closed box equipped with camera where patients could fit their hands) or by the hospital photographer. In both cases, capturing conditions were standardized: both hands facing up/down, fixed background (green for the device and grey for the photographer), controlled lighting and zoom levels. An aspect that was not standardized was the portion of the wrists to be included as the imaging focus was the hands. Pictures were anonymized by the removal of all patient-identifying information.

2.1 | Hand eczema data set

The HE data set was composed of 312 high-resolution pictures (156 front and back hands pairs) annotated by a team of 11 experienced dermatologists for eczema lesions, healthy skin and background. When annotations for the same picture were available, the majority consensus was computed. The data set was randomly split into 249 pictures for training and 63 for testing, ensuring no leak of pictures from the same patient. To leverage the full pixel resolution, all pictures were divided into square patches of size 512 pixels resulting in 7755 training patches and 1937 test patches.

2.2 | Hand eczema DLM training

The HE DLM was based on the U-Net[14] architecture with a ResNet[15] backbone pretrained on ImageNet.[16] HE training patches were resized to squares of 256 pixels size and the DLM was trained for 40 epochs, with a batch size of 16, the Adam[17] optimizer and one cycle scheduling[18] for a learning rate initialized at 1 e-4. To mitigate data set imbalance, we used a combination of the dice loss[19] and the focal loss.[20] Data augmentation operations consisted in random rotations, flips, brightness, contrast, perspective and zoom changes.

2.3 | Hand anatomy data set

The anatomy data set comprised 215 high-resolution hand pictures with 99 front hands and 116 back hands. Each picture was annotated by one medical student with 37 anatomical regions presented in Figure 1, including the wrist and "non-hand" (anything else) regions. The correspondence between these anatomical regions and the HECSI regions is presented in the Table S1. The data set was randomly divided into 171 pictures for training and 44 pictures for testing performance, ensuring no leak.

2.4 | Anatomy DLM training

The architecture of the anatomy DLM was similar to the HE DLM. We used the same training conditions except that the anatomy

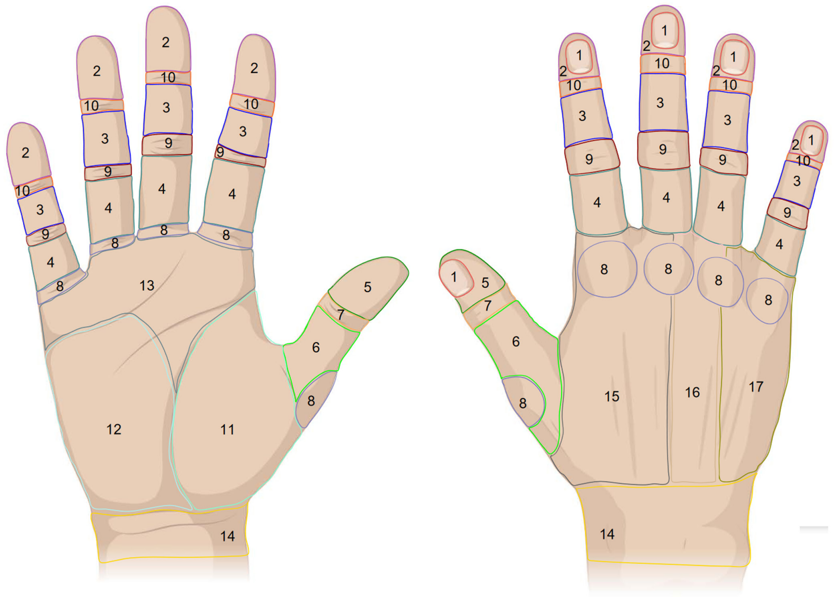


FIGURE 1 Hands' anatomical regions. his schema presents the different hands' anatomical regions used in this work: nail (1), fingers II-V distal (2), fingers II-V middle (3), fingers II-V proximal (4), thumb distal (5), thumb proximal (6), interphalangeal (IP) joint I (7), metacarpophalangeal (MCP) I-V (8), proximal IP (PIP) II-V (9), distal IP (DIP) II-V (10), thenar (11), hypothenar (12), palm (13), wrist (14), dorsal radial (15), dorsal middle (16) and dorsal lateral (17).

training pictures were resized to squares of 380 pixels side-size and that the batch size was fixed to 4.

2.5 | Hand eczema assessment workflow

The workflow of our HE severity assessment system (Figure 2) essentially consists of five steps. First, the patient's hands are photographed from both sides. Then, the HE DLM predicts the eczema lesions in the pictures, followed by the mapping of the anatomical regions by the anatomy DLM (these two steps could be executed in parallel). Finally, the predictions are merged and a disease report is generated, providing a textual description of the disease together with a quantification of eczema surface per anatomical regions.

2.6 | Analysis

The performance of the HE and anatomy DLMs were evaluated on the respective test data sets using the precision and sensitivity metrics with 95% confidence interval (CI). The CI were determined using the non-parametric bootstrap resampling method. In the case of the HE DLM, the full picture predictions were first reconstructed from the individual test patches predictions before computing the performance metrics. Furthermore, we evaluated the intraclass correlation (ICC) of the predicted HE surface with experts' annotations.

We also analysed the performance of both DLMs after aggregating their predictions over the HECSI anatomical regions. In the case of the anatomy data set, we could merge the anatomical regions labelled by the student into HECSI regions (as per Table S1), while for the HE DLM, we used the HECSI regions obtained from the anatomy DLM predictions.

To gain insights on the HE data set, we computed the average eczema surface per anatomical region with standard deviation and median. This analysis was performed based on the anatomy DLM predictions of the full HE data set and the dermatologists' HE labels.

Finally, taking an example patient case from the HE test set, we automatically generated a textual disease report with corresponding eczema anatomical stratification tables.

3 | RESULTS

3.1 | Hand eczema

The performance of the HE DLM was evaluated on the HE test set pictures (Table 1). When evaluating the performance over the full pictures, the DLM achieved a precision of 75% (CI 64–82) and a sensitivity of 69% (CI 55–81). The ICC of the predicted HE surface was 0.94 (CI 0.90–0.96) indicating a very strong correlation with experts' annotations.

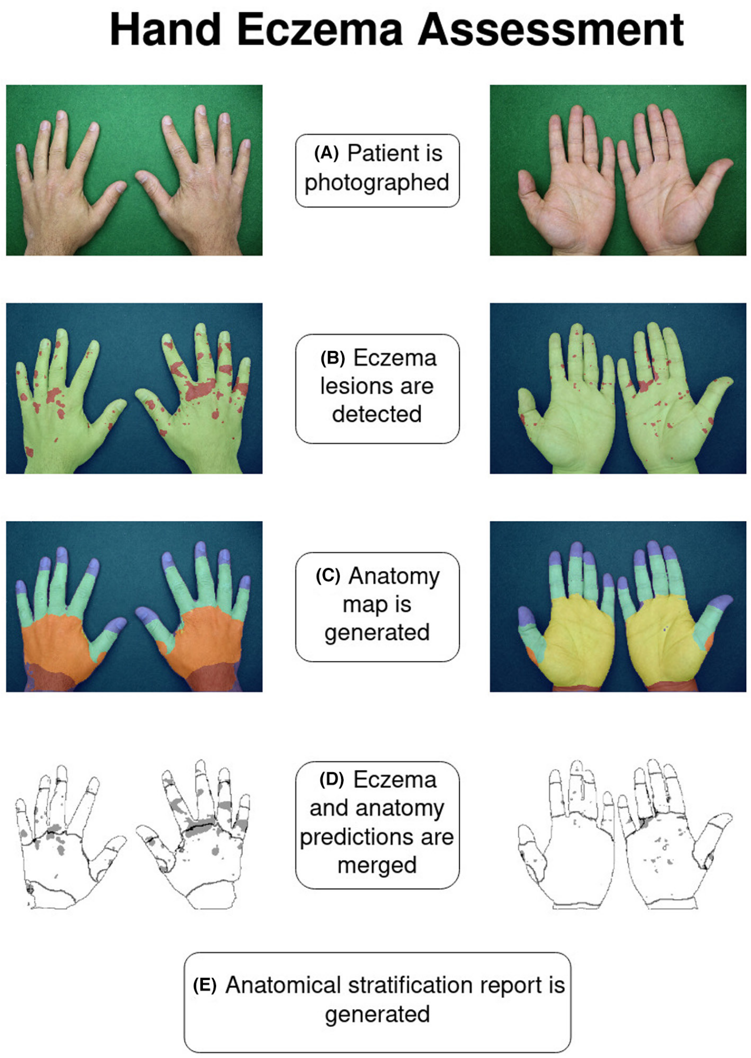

Hand Eczema Assessment
(A) Patient is photographed
(B) Eczema lesions are detected
(C) Anatomy map is generated
(D) Eczema and anatomy predictions are merged
(E) Anatomical stratification report is generated

FIGURE 2 Hand eczema assessment workflow. This figure presents a patient's front and back hand pictures (A), the corresponding hand eczema deep learning model (DLM) predictions (B), the hands anatomical regions (aggregated over the same regions assessed in the hand eczema severity index system for visual clarity) mapped by the anatomy DLM (C) and the combination of both DLMs predictions (D). In (B), the background is violet, the skin is green and the eczema lesions are red. In (C), the non-hand region is violet, the wrist is red, the palm of the hand is yellow, the fingers (without tips) is light blue, the fingertips are dark blue and the back of hand is orange.

Considering the HECSI regions (predicted by the anatomy DLM) separately, we observed that the HE DLM was more precise but less sensitive on the palm of hands, fingers and fingertips similar to the average performance on full pictures. However, the opposite occurred for the wrist and back of hands, both of which tended to be covered by hairs, a known source of confusion for segmentation approaches in such settings.

The analysis of eczema anatomical stratification of the HE data set (for HECSI regions in Table 2 and for all anatomical regions in Table S2) revealed, that the regions mostly covered by eczema lesions were the fingers and fingertips with 13.1% and 12%, respectively, followed by palm of hands with 11.6%. The wrist and back of hands had the least coverage with an average of 4.7% and 5.8% and a median close to 0%. Thus, more than half of the pictures did not have any eczema lesions on these regions, which explains the relatively large confidence intervals of the predictions. For all regions, the eczema surface standard deviation was high, above 15%.

3.2 | Hand anatomy

The performance of the anatomy DLM was evaluated on the anatomy test set pictures (Table 3). In average the DLM achieved a precision of 83% (CI 80–85) and a sensitivity of 85% (CI 82–88). The limits of wrists with arms were challenging to determine due to the lack of standardization of this particular region in the training pictures. The DLM also had difficulties for some of the MPCs (especially MPC1 on the thumb) and DIPs regions, because of their small size and unclear boundaries with respect to the other anatomical regions.

For the combined experiment using both the anatomy and HE DLMs, the regions were aggregated over the HECSI regions. This

TABLE 1 Performance of the eczema deep learning model.

Regions	Category	Precision	Sensitivity
Full pictures	Background	100% (100–100)	100% (100–100)
	Skin	95% (92–98)	97% (96–98)
	Eczema	75% (64–82)	69% (55–81)
Fingertips	Eczema	74% (65–79)	70% (63–77)
Fingers (without tips)	Eczema	78% (68–84)	69% (59–79)
Palm of hand	Eczema	78% (64–86)	84% (69–90)
Back of hand	Eczema	66% (23–85)	50% (20–85)
Wrist	Eczema	68% (27–87)	44% (19–86)
Average of HECSI regions	Eczema	71% (53–80)	62% (50–78)

Note: Performance evaluated on the hand eczema test set by comparing the eczema deep learning model predictions with the dermatologists' lesion annotations. Parentheses indicate the 95% confidence interval. The hand eczema severity index (HECSI) regions were predicted by the anatomy deep learning model.

TABLE 2 Anatomical stratification of eczema lesions.

Regions	Surface average	Surface standard deviation	Surface median	Surface interquartile range
Back of hand	5.8%	17.4%	0.2%	1.9%
Fingertips	12%	19.2%	4%	13.3%
Fingers (without tips)	13.1%	21.8%	3.9%	12.8%
Palm of hand	11.6%	22.8%	1.6%	10.7%
Wrist	4.7%	16.5%	0%	0%

Note: Eczema surface repartition over the hand eczema severity index anatomical regions. Evaluated on the full hand eczema data set using dermatologists' lesion annotations and the anatomy deep learning model predictions.

yielded a high performance since the regions' separations are more clearly defined: the average precision and sensitivity were 91% (CI 90–92) and 94% (CI 93–94).

3.3 | Disease report generation

Figure 1 presents a random patient case from the HE test data set with the predicted eczema lesions and HECSI anatomical regions. Our system automatically generated the following textual description for this patient's condition: "The patient's hands show eczema lesions on both the palmar and back sides, namely on 4.8% of the fingertips, 11% of the fingers (without tips), 1.5% of the palms, 3% of the back of hands and 1.1% of the wrists".

4 | DISCUSSION

Hand eczema is a highly prevalent disease that is often chronic and requires diligent and detailed clinical follow-up. Objective disease quantification is key for judging the success of clinical management but is challenging to perform in practice, as it requires time and expertize. In this work, we present an automated method to analyse the anatomical repartition of HE lesions from patients' hands pictures. Our approach leveraged two DLMs, one to segment HE lesions with precision and sensitivity 75% (CI 64–82) and 69% (CI 55–81), the second to segment hands anatomical regions with precision and sensitivity 83% (CI 80–85) and 85% (CI 82–88). In application of our approach, we could automatically generate the quantitative and textual description of a test patient's condition as well as compute statistics on the anatomical repartition of eczema lesions in our data set.

Commenting on the reported model performance, the sensitivity of a DLM is always a trade-off with its precision. The large confidence intervals are explained by the small size of the test data set together with the observation that a large proportion of the pictures had little to no eczema in certain anatomical regions. Given additional training data, the model sensitivity and precision could theoretically be improved. It is important to consider that the perfect segmentation of eczema lesions is not the most important objective of this study but rather the robust quantification of eczema lesions in a reproducible manner to enable precise disease monitoring in time and patient follow-up.

To the best of our knowledge, this study is the first to generate a mapping of hands' anatomical regions from patients' pictures as well as the anatomical stratification of HE lesions. Other work related to hand segmentation focused either on hand detection,[21] palm region extraction for biometrics,[22] gesture recognition[23] or bone segmentation from ultrasound and MRI scans.[24,25] Previous work on automated eczema severity assessment were based on smaller data sets and mainly proposed lesion segmentation approaches,[26] some with classification of the overall severity level.[27-29] One study's approach consisted in the detection (as opposed to segmentation) of atopic eczema lesions based on 1393 patients' pictures followed by the severity classification of seven clinical signs.[30] Segmentation and classification of eczema lesions was also performed on histopathological slides.[31]

TABLE 3 Performance of the hand anatomy deep learning model.

Regions	Precision	Sensitivity
Non-hand	99% (99–99)	97% (97–98)
DIP2	71% (58–79)	82% (72–88)
DIP3	77% (72–81)	84% (74–90)
DIP4	72% (67–78)	84% (73–90)
DIP5	75% (69–80)	85% (80–90)
IP	79% (76–82)	84% (81–87)
MCP1	64% (57–71)	79% (74–84)
MCP2	74% (69–79)	82% (74–86)
MCP3	75% (69–79)	84% (79–88)
MCP4	68% (60–75)	77% (69–83)
MCP5	72% (65–77)	79% (75–84)
PIP2	84% (75–90)	88% (82–92)
PIP3	87% (84–90)	85% (72–91)
PIP4	84% (78–88)	87% (84–90)
PIP5	84% (79–87)	86% (82–89)
Dorsal mid	72% (67–77)	76% (69–81)
Dorsal radial	86% (81–89)	85% (82–88)
Dorsal ulnar	87% (85–89)	77% (69–81)
Hypothenar	87% (84–90)	89% (81–95)
Index distal	85% (78–92)	88% (82–92)
Index middle	84% (74–91)	88% (83–92)
Index proximal	87% (81–92)	89% (83–93)
Little f. distal	90% (87–93)	89% (82–93)
Little f. middle	91% (89–93)	85% (82–88)
Little f. proximal	87% (85–90)	88% (86–91)
Middle f. distal	91% (87–94)	87% (80–93)
Middle f. middle	92% (87–94)	88% (82–92)
Middle f. proximal	89% (85–92)	88% (80–92)
Nail	89% (86–91)	83% (78–86)
Palm	89% (86–93)	86% (84–89)
Ring f. distal	87% (82–92)	87% (76–93)
Ring f. middle	89% (83–94)	86% (78–91)
Ring f. proximal	88% (84–91)	88% (84–91)
Thenar	88% (83–91)	89% (85–92)
Thumb distal	92% (90–93)	89% (86–92)
Thumb proximal	87% (83–90)	80% (76–83)
Wrist	69% (64–74)	86% (83–89)
Average	83% (80–85)	85% (82–88)
HECSI regions	Precision	Sensitivity
Non-hand	99% (99–99)	97% (97–98)
Fingertips	96% (95–96)	94% (92–95)
Fingers (without tips)	94% (93–95)	94% (93–95)
Palm of hand	96% (95–97)	98% (96–98)
Back of hand	93% (90–94)	93% (92–95)
Wrist	69% (64–74)	86% (83–89)
Average	91% (90–92)	94% (93–94)

Note: Performance evaluated on the anatomy test set by comparing the anatomy deep learning model predictions with the medical student's annotations. Parentheses indicate the 95% confidence interval. HECSI stands for the hand eczema severity index, IP for interphalangeal joint I, MCP for metacarpophalangeal, PIP for proximal IP, DIP for distal IP, f for finger.

A particular challenge faced in this study concerned the boundaries of the different hand anatomical regions. These are not clearly defined in the anatomy literature and are subject to personal interpretation in practice. In this work, unclear region frontiers were clarified with a board-certified dermatologist. The difficulties of the anatomy DLM with the determination of wrists' limits on arms were caused by variations in the training set pictures of the visible portion of wrists. This aspect was not fully standardized in the collection protocol as the photographer's goal was to capture full hands.

Further clinical studies are required to robustly differentiate mild, moderate and severe HE. Our method can be used to support clinicians in this regard by providing precise quantification of the anatomical repartition of eczema surface. These estimates have the advantage to be automated and reproducible, independent from experience or training, eliminating inter- and intra-observer variance. The results can be automatically translated to disease reports and thus assist in the documentation of patients' conditions. This approach enables less experienced clinicians to produce objective and comparable evaluation of their patients. Follow-ups can be performed remotely, either by directly integrating DLMs into mobile phone apps or by serving predictions via a web server. In this case, the picture acquisition process should be guided to ensure the captured pictures are sufficiently standardized and similar to this study's data sets. When HECSI scores are to be computed, predicted surface estimates can be combined with dermatologist's manual severity grading of HE clinical signs, all of which can be achieved remotely with classic store-and-forward teledermatology.[32]

With our method, the typical anatomical stratification of eczema lesions could be evaluated from large HE databases (similarly to Table 2 and Table S2) to help determine the regions that are more prone to develop eczema lesions and to which proportions. Similarly, the clinical evolution of individual patients' HE, and the effects of treatment could be monitored with high precision and benefit drug development efforts.

The presented hand anatomy DLM is not restricted to HE and can be equivalently used to determine the anatomical repartition of other diseases affecting hands such as lentigo, psoriasis, vitiligo or palmoplantar pustulosis. Furthermore, our anatomical segmentation approach can be applied equivalently to other body regions enabling similar applications.

4.1 | Limitations

One limitation of this study was caused by the data sets' characteristics, which only comprised hands from skin type 1 to 3 on the Fitzpatrick scale photographed in a standardized position (cf. Figure 1). As a result, the DLMs presented in this study will underperform on pictures from patient with other skin types or with hands in different position, for example, closed fists. Furthermore, the DLM could mistakenly segment benign skin lesions such as seborrheic keratoses since they were not included in the training data set. These issues can be mitigated by retraining the DLMs on more complete data sets. Another limitation by design is that our approach does not evaluate the severity of eczema clinical signs, necessary to fully automate the HECSI score. This choice was caused by the lack of necessary data (each feature is ranked on four severity levels, all of which would require corresponding pictures to train a DLM for automation) and is planned as future work together with a prospective study on how HECSI scores correlate with this study's surface predictions. Finally, picture-based approaches such as ours, must inevitably base their predictions on limited information. Thus, for applications with high precision requirements, it is of interest to explore other image modalities that provide additional information such as multispectral imaging.[33]

5 | CONCLUSION

Taken together, by quantifying aspects of patients' conditions, our approach translates information that could so far, only be inferred and interpreted by dermatologists, into an easily shareable, objective and accessible digest. The determination of condition-specific actionable rules is the next step to empower less specialized clinicians and scale-up HE care.

AUTHOR CONTRIBUTIONS

Ludovic Amruthalingam: Conceptualization, data curation, formal analysis, investigation, methodology, project administration, software, validation, visualization, writing—original draft and writing—review and editing. Nora Mang: Data curation. Philippe Gottfrois: Methodology and validation. Alvaro Gonzalez Jimenez: Methodology and validation. Julia-Tatjana Maul: Methodology, validation and writing—review and editing. Michael Kunz: Data curation, methodology and validation. Marc Pouly: Conceptualization, funding acquisition, methodology, project administration, resources, supervision, validation and writing—review and editing. Alexander A. Navarini: Conceptualization, data curation, funding acquisition, methodology, project administration, resources, supervision, validation and writing—review and editing.

FUNDING INFORMATION

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

CONFLICT OF INTEREST

Maul JT has served as advisor and/or received speaking fees and/or participated in clinical trials sponsored by AbbVie, Almirall, Amgen, BMS, Celgene, Eli Lilly, LEO Pharma, Janssen-Cilag, MSD, Novartis, Pfizer, Pierre Fabre, Roche, Sanofi and UCB. Navarini AA 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.

DATA AVAILABILITY STATEMENT

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

ORCID

Ludovic Amruthalingam https://orcid.org/0000-0001-5980-5469
Marc Pouly https://orcid.org/0000-0002-9520-4799

REFERENCES

1. Agner T, Aalto-Korte K, Andersen KE, et al. European environmental and contact dermatitis research group. Classification of hand eczema. *J Eur Acad Dermatol Venereol*. 2015;29(12):2417-2422.
2. Agner T, Elsner P. Hand eczema: epidemiology, prognosis and prevention. *J Eur Acad Dermatol Venereol*. 2020;34:4-12.
3. Nørreslet LB, Agner T, Sørensen JA, Ebbehøj NE, Bonde JP, Fisker MH. Impact of hand eczema on quality of life: metropolitan versus non-metropolitan areas. *Contact Dermatitis*. 2018;78(5):348-354.
4. Oosterhaven JA, Ofenloch RF, Schuttelaar ML. Validation of the Dutch quality of life in hand eczema questionnaire (QOLHEQ). *British Journal of Dermatology*. 2020;183(1):86-95.
5. Politiek K, Oosterhaven JA, Vermeulen KM, Schuttelaar ML. Systematic review of cost-of-illness studies in hand eczema. *Contact Dermatitis*. 2016;75(2):67-76.
6. Apfelbacher CJ, Ofenloch RF, Weisshaar E, et al. Chronic hand eczema in Germany: 5-year follow-up data from the CARPE registry. *Contact Dermatitis*. 2019;80(1):45-53.
7. De León FJ, Berbegal L, Silvestre JF. Management of chronic hand eczema. *Actas Dermosifiliogr (English Edition)*. 2015;106(7):533-544.
8. Weistenhöfer W, Baumeister T, Drexler H, Kütting B. An overview of skin scores used for quantifying hand eczema: a critical update according to the criteria of evidence-based medicine. *Br J Dermatol*. 2010;162(2):239-250.
9. Held E, Skoet R, Johansen JD, Agner T. The hand eczema severity index (HECSI): a scoring system for clinical assessment of hand eczema. A study of inter-and intraobserver reliability. *Br J Dermatol*. 2005;152(2):302-307.
10. Oosterhaven JA, Schuttelaar ML. Responsiveness and interpretability of the hand eczema severity index. *Br J Dermatol*. 2020;182(4):932-939.
11. Yüksel YT, Agner T, Ofenloch R. New evidence on the minimal important change (MIC) for the hand eczema severity index (HECSI). *Contact Dermatitis*. 2021;85(2):164-170.
12. Yuan Y, Chen X, Chen X, Wang J. Segmentation transformer: object-contextual representations for semantic segmentation. arXiv preprint arXiv:1909.11065. 2019.
13. Wortsman M, Ilharco G, Gadre SY, et al. Model soups: averaging weights of multiple fine-tuned models improves accuracy without increasing inference time. arXiv preprint arXiv:2203.05482. 2022.
14. Ronneberger O, Fischer P, Brox T. U-net: convolutional networks for biomedical image segmentation. *International Conference on Medical Image Computing and Computer-Assisted Intervention*. Springer; 2015:234-241.
15. He K, Zhang X, Ren S, Sun J. Deep residual learning for image recognition. *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. IEEE; 2016:770-778.
16. Deng J, Dong W, Socher R, Li LJ, Li K, Fei-Fei L. Imagenet: a large-scale hierarchical image database. *2009 IEEE Conference on Computer Vision and Pattern Recognition*. Ieee; 2009:248-255.
17. Kingma DP, Ba J. Adam: a method for stochastic optimization. arXiv preprint arXiv:1412.6980. 2014.
18. 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 .

19. 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(2):576-589.
20. Lin TY, Goyal P, Girshick R, He K, Dollár P. Focal loss for dense object detection. *Proceedings of the IEEE International Conference on Computer Vision*. IEEE; 2017:2980-2988.
21. Urooj A, Borji A. Analysis of hand segmentation in the wild. *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. IEEE; 2018:4710-4719.
22. Ito K, Suzuki Y, Kawai H, et al. HandSegNet: hand segmentation using convolutional neural network for contactless palmprint recognition. *IET Biom*. 2022;11(2):109-123.
23. Paul S, Bhattacharyya A, Mollah AF, Basu S, Nasipuri M. Hand segmentation from complex background for gesture recognition. *Emerging Technology in Modelling and Graphics*. Springer; 2020:775-782.
24. Kayalibay B, Jensen G, van der Smagt P. CNN-based segmentation of medical imaging data. arXiv preprint arXiv:1701.03056. 2017.
25. Sultan MS, Martins N, Ferreira MJ, Coimbra MT. Segmentation of bones and MCP joint region of the hand from ultrasound images. *2015 37th Annual International Conference of the IEEE Engineering in Medicine and Biology Society (EMBC)*. IEEE; 2015:3001-3004.
26. Schnürle S, Pouly M, vor der Brück T, Navarini A, Koller T. On using support vector machines for the detection and quantification of hand eczema. *ICAART*. Springer; 2017:75-84.
27. Alam MN, Munia TT, Tavakolian K, Vasefi F, MacKinnon N, Fazel-Rezai R. Automatic detection and severity measurement of eczema using image processing. *2016 38th Annual International Conference of the IEEE Engineering in Medicine and Biology Society (EMBC)*. IEEE; 2016:1365-1368.
28. Nisar H, Ch'ng YK, Ho YK. Automatic segmentation and classification of eczema skin lesions using supervised learning. *2020 IEEE Conference on Open Systems (ICOS)*. IEEE; 2020:25-30.
29. Ch'ng YK, Nisar H, Yap VV, Yeap KH, Tang JJ. Segmentation and grading of eczema skin lesions. *2014 8th International Conference on Signal Processing and Communication Systems (ICSPCS)*. IEEE; 2014:1-5.
30. Pan K, Hurault G, Arulkumaran K, Williams HC, Tanaka RJ. EczemaNet: automating detection and severity assessment of atopic dermatitis. *International Workshop on Machine Learning in Medical Imaging*. Springer; 2020:220-230.
31. Scheurer J, Ferrari C, Berenguer Todo Bom L, Beer M, Kempf W, Haug L. Semantic segmentation of histopathological slides for the classification of cutaneous lymphoma and eczema. *Annual Conference on Medical Image Understanding and Analysis*. Springer; 2020:26-42.
32. Kanthraj GR. Classification and design of teledermatology practice: what dermatoses? Which technology to apply? *J Eur Acad Dermatol Venereol*. 2009;23(8):865-875.
33. Hald M, Thyssen JP, Zachariae C, et al. Multispectral imaging of hand eczema. *Contact Dermatitis*. 2019;81(6):438-445.

SUPPORTING INFORMATION

Additional supporting information can be found online in the Supporting Information section at the end of this article.

Appendix S1:

How to cite this article: Amruthalingam L, Mang N, Gottfrois P, et al. Objective hand eczema severity assessment with automated lesion anatomical stratification. *Exp Dermatol*. 2023;00:1-8. doi:10.1111/exd.14744

SUPPORTING INFORMATION

In these supplementary materials we provide additional information and results for the following:

- Mapping of the hand’s anatomical regions
- Patient sample detailed anatomy predictions
- Confusion matrix of the HE DLM
- Detailed anatomical stratification of HE lesions in the full HE dataset
- Extended results on disease report generation
- Robustness of the DLMs performance: randomness and data augmentation

Mapping of the hand's anatomical regions

Table S1 HECSI regions correspondance with anatomical regions

HECSI region	Anatomical region
Non-hand	Non-hand
Fingertips	Index distal
	Little f. distal
	Middle f. distal
	Nail
	Ring f. distal
	Thumb distal
Fingers (without tips)	DIP2
	DIP3
	DIP4
	DIP5
	IP
	PIP2
	PIP3
	PIP4
	PIP5
	Index middle
	Index proximal
	Little f. middle
	Little f. proximal
	Middle f. middle
	Middle f. proximal
	Ring f. middle
	Ring f. proximal
	Thumb proximal
Palm of hand	Hypothenar
	Palm
	Thenar
	MCP2
	MCP3
	MCP4
	MCP5
Back of hand	Dorsal mid
	Dorsal radial
	Dorsal ulnar
	MCP1
	MCP2
	MCP3
	MCP4
	MCP5
Wrist	Wrist

Patient sample detailed anatomy predictions

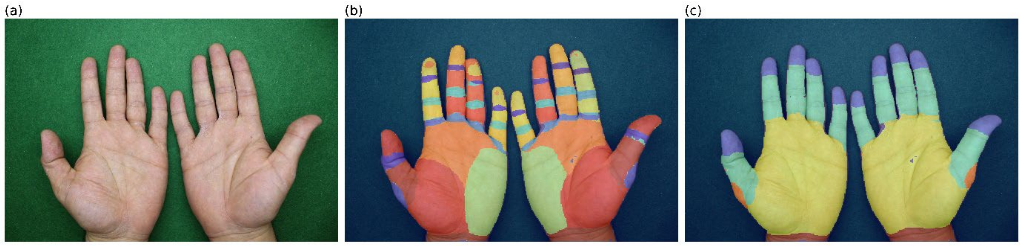


Figure S1 Patient's front hands anatomy predictions

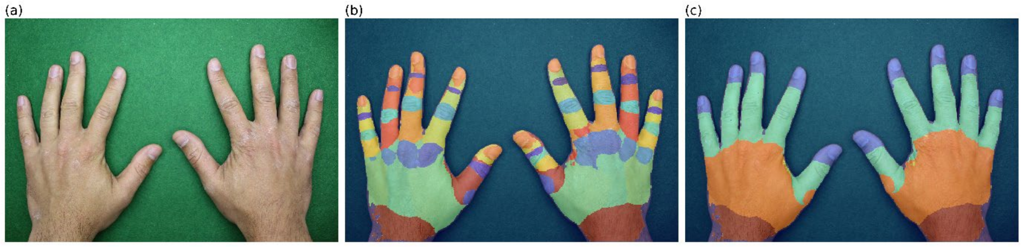


Figure S2 Patient's back hands anatomy predictions

In both Figures S1 and S2, the sample pictures are shown in (a), anatomical predictions in (b) and the aggregated HECSI regions in (c).

Confusion matrix of the HE DLM

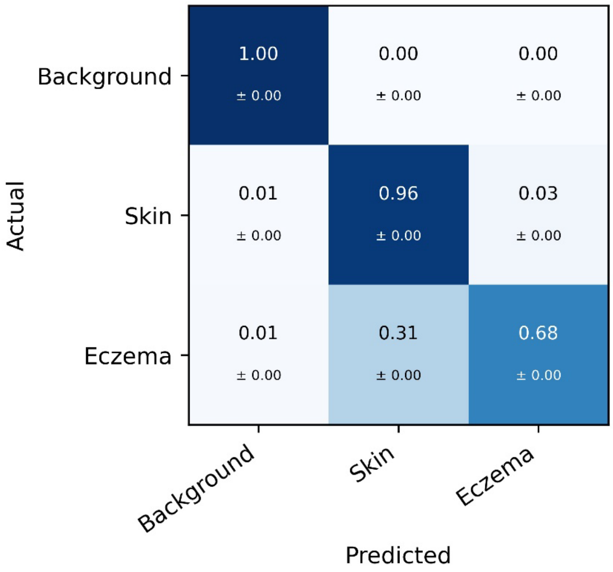


Figure S3 HE DLM confusion matrix

The confusion matrix shows the average proportion of pixels classified between the background, skin and eczema classes. The vertical axis represents the true pixel labels, while the horizontal axis shows the predicted labels.

Detailed anatomical stratification of HE lesions in the full HE dataset

Table S2 HE lesions' surface repartition on the hand's anatomical regions

Region	Average surface	STD surface	Median surface	IQR surface
Non-hand	0.2%	0.4%	0.1%	0.2%
DIP2	19.4%	30.6%	0%	29.5%
DIP3	18.3%	30.7%	0%	24.4%
DIP4	16.4%	28.5%	0.1%	17.6%
DIP5	16%	27.3%	0.9%	17.4%
IP	13.8%	24.9%	0.3%	16.2%
MCP1	6.6%	20.4%	0%	0%
MCP2	11.8%	25.3%	0%	6.4%
MCP3	11.3%	25%	0%	7.7%
MCP4	10.8%	24.3%	0%	7.5%
MCP5	12.6%	26%	0%	11.7%
PIP2	16%	28%	0%	18.4%
PIP3	13.2%	26.2%	0%	10.2%
PIP4	12.1%	24.1%	0.4%	9.1%
PIP5	15.1%	27.1%	1.6%	14.3%
Dorsal mid	4.1%	17.6%	0%	0%
Dorsal radial	4.7%	17.9%	0%	0.3%
Dorsal ulnar	5.4%	19.4%	0%	0%
Hypothenar	11.7%	24.3%	0.8%	8.1%
Index distal	16.4%	25.6%	1.6%	24.2%
Index middle	15.5%	27.4%	0%	18.7%
Index proximal	12%	25%	0%	8.9%
Little f. distal	11.7%	20.9%	2%	12.6%
Little f. middle	14.6%	26.2%	1.3%	14.7%
Little f. proximal	10.9%	23.8%	0%	7.2%
Middle f. distal	14.9%	26.6%	0%	15.6%
Middle f. middle	14.9%	28.5%	0%	12.6%
Middle f. proximal	10.1%	23.9%	0%	5%
Nail	8.3%	19.6%	0.1%	5.1%
Palm	10.8%	23.5%	0.6%	7.4%
Ring f. distal	12%	22.3%	0.8%	11.4%
Ring f. middle	13%	25.1%	0.5%	11.2%
Ring f. proximal	9.6%	22.2%	0%	5.5%
Thenar	11.4%	23.9%	0.7%	8.5%
Thumb distal	14.9%	26.5%	0%	18.5%
Thumb proximal	11.3%	22.2%	0.8%	10.6%
Wrist	4.7%	16.5%	0%	0%

Eczema coverage of the hand anatomical regions. Evaluated on the full HE dataset and based on the DLM anatomical predictions.

Extended results on disease report generation

Here we detail the predicted eczema anatomical stratification for the patient presented in the paper.

Table S3 Anatomical stratification over HECSI regions of patient's HE lesions

Region	Predicted surface
Fingertips	4.8%
Fingers (without tips)	11%
Palm of hand	1.5%
Back of hand	9%
Wrist	1.1%

Table S4 Detailed anatomical stratification of patient's HE lesions

Region	Predicted surface
Non-hand	0.3%
DIP2	4.6%
DIP3	6.5%
DIP4	18.8%
DIP5	9.6%
IP	4.7%
MCP1	43.5%
MCP2	3.6%
MCP3	34.4%
MCP4	23.3%
MCP5	10.4%
PIP2	20.7%
PIP3	4.2%
PIP4	24.2%
PIP5	28.4%
Dorsal mid	5.7%
Dorsal radial	1.0%
Dorsal ulnar	2.0%
Hypothenar	0.9%
Index distal	12.6%
Index middle	6.2%
Index proximal	6.9%
Little f. distal	5.4%
Little f. middle	5.3%
Little f. proximal	11.8%
Middle f. distal	4.9%
Middle f. middle	10.3%
Middle f. proximal	25.6%
Nail	3.5%
Palm	3.4%
Ring f. distal	6.9%
Ring f. middle	4.7%
Ring f. proximal	8.4%
Thenar	0.6%

Thumb distal	1.7%
Thumb proximal	4.0%
Wrist	1.1%

Robustness of the DLMs performance: randomness and data augmentation

The performance of DLMs can be influenced by the random initialization of their parameters, especially when training datasets are small with respect to DL standards. To quantify the randomness of the reported performance, we trained each DLM with twenty different random seeds and evaluated the mean, median and standard deviation (std) of the achieved precision and sensitivity (Table S5). For the HE DLM, we measured a precision std of 1.17% and a sensitivity std of 1.25% while for the anatomy DLM the std amounted to 0.22% for the precision and 0.18% for the sensitivity.

Table S5 Evaluation of randomness in DLMs' performance

	HE DLM		**Anatomy DLM**	
	Precision	**Sensitivity**	**Precision**	**Sensitivity**
Mean	74.7%	67.3%	83.5%	84.7%
STD	1.17%	1.25%	0.22%	0.18%
Median	74.65%	67.47%	83.53%	84.76%

To increase the DLMs' generalization capability, we applied random rotations, flips, contrast, brightness, perspective and zoom augmentations during training. An interesting experiment illustrating how well the DLMs handle transformed images is to evaluate their test performance after applying the different transformations separately. We performed this analysis based on twenty different random seeds and calculated the mean, median, std of the precision and sensitivity for both DLMs (Table S6).

Table S6 Evaluation of DLMs' performance robustness against data augmentation

		HE DLM		**Anatomy DLM**	
Transforms		**Precision**	**Sensitivity**	**Precision**	**Sensitivity**
Original	Mean	75.1%	68.6%	83.1%	85.4%
	STD	0.0%	0.0%	0.0%	0.0%
	Median	75.1%	68.6%	83.1%	85.4%
Rotation	Mean	76.6%	66.1%	79.2%	81.4%
	STD	0.5%	0.6%	0.6%	0.6%
	Median	76.6%	66.1%	79.0%	81.2%
Flip	Mean	74.7%	67.0%	80.9%	83.6%
	STD	0.7%	0.2%	0.3%	0.2%
	Median	74.9%	67.0%	81.0%	83.5%
Brightness and contrast	Mean	74.5%	67.6%	77.6%	76.9%
	STD	1.0%	0.8%	1.7%	2.7%
	Median	74.5%	67.8%	78.0%	77.2%
Perspective	Mean	76.1%	70.2%	80.5%	81.1%
	STD	0.5%	1.3%	0.0%	0.0%
	Median	76.5%	71.0%	80.5%	81.1%
Zoom	Mean	76.3%	67.4%	83.1%	85.0%
	STD	0.4%	0.4%	0.1%	0.2%
	Median	76.3%	67.5%	83.1%	85.0%

Chapter 8

Generation of Synthetic Dermatology Images

One of the main challenges faced in the development of deep learning applications for dermatology is the availability of data, as we illustrated in section 4.3.2. Public datasets are necessary to reproduce and compare published research results. However, this is often not possible due to legal constraints or commercial interests around medical data. The generation of synthetic data is a promising approach to create publicly shareable datasets, even when the original available data is sensitive and must remain private. In section 8.1, we present our conference paper on the generation of artificial dermatology images.

8.1 Applications of Generative Adversarial Networks to Dermatologic Imaging

This conference paper was accepted [58] at the 2020 International Association of Pattern Recognition Workshop on Artificial Neural Networks in Pattern Recognition [1]. We hypothesized that synthetic skin lesion images could be produced using generative adversarial networks. In this work, we validate this hypothesis and generate artificial data for the two main dermatology image modalities, namely photography of eczema lesions and dermoscopy imaging of melanocytic lesions. After verification for patient identifying features, this synthetic data could be shared to establish standard evaluation datasets. It could also be used to improve performance of deep learning models by training on a combination of original and synthetic data.

[1] Full text via DOI: `https://doi.org/10.1007/978-3-030-58309-5_15` (Accessed: 2nd February 2023)

Applications of Generative Adversarial Networks to Dermatologic Imaging

Fabian Furger[1], Ludovic Amruthalingam[2]( ), Alexander Navarini[2,3], and Marc Pouly[1]

[1] Department of Information Technology, Lucerne University of Applied Sciences and Arts, Lucerne, Switzerland
furgerfabian@hotmail.com, marc.pouly@hslu.ch

[2] Department of Biomedial Engineering, University of Basel, Basel, Switzerland
ludovic.amruthalingam@unibas.ch

[3] Department of Dermatology, University Hospital of Basel, Basel, Switzerland
alexander.navarini@usb.ch

Abstract. While standard dermatological images are relatively easy to take, the availability and public release of such data sets for machine learning is notoriously limited due to medical data legal constraints, availability of field experts for annotation, numerous and sometimes rare diseases, large variance of skin pigmentation or the presence of identifying factors such as fingerprints or tattoos. With these generic issues in mind, we explore the application of Generative Adversarial Networks (GANs) to three different types of images showing full hands, skin lesions, and varying degrees of eczema. A first model generates realistic images of all three types with a focus on the technical application of data augmentation. A perceptual study conducted with laypeople confirms that generated skin images cannot be distinguished from real data. Next, we propose models to add eczema lesions to healthy skin, respectively to remove eczema from patient skin using segmentation masks in a supervised learning setting. Such models allow to leverage existing unrelated skin pictures and enable non-technical applications, e.g. in aesthetic dermatology. Finally, we combine both models for eczema addition and removal in an entirely unsupervised process based on CycleGAN without relying on ground truth annotations anymore. The source code of our experiments is available on https://github.com/furgerf/GAN-for-dermatologic-imaging.

Keywords: Generative Adversarial Networks · Dermatology

1 Introduction

Generative Adversarial Networks (GANs), initially proposed in [8] have since then produced impressive results in a variety of synthetic data generation tasks. In contrast to other deep learning methods, which are notoriously data-intensive, GANs achieve good results even with relatively small data sets [2,7]. This makes GANs attractive for domains where training data is difficult or expensive to obtain. A standard example is the medical field, where specialized machinery

F.-P. Schilling and T. Stadelmann (Eds.): ANNPR 2020, LNAI 12294, pp. 187–199, 2020.
https://doi.org/10.1007/978-3-030-58309-5_15

may be needed or occurrences of pathologies may be hard to find. Using data sets augmented with GAN-generated synthetic data to train machine learning models has improved performance in a variety of medical domains [3,9,12].

Dermatology is one domain particularly suited for the application of deep learning models, but with far too few publicly-available data sets compared to the diversity of the cases encountered in clinical practice. Therefore, the idea to leverage the GAN framework to generate new samples is very promising. However, applications in dermatology are to this date still rare. One example is *MelanoGAN* [2], which generates images of skin lesions from ISIC 2017 [5]. The authors compare the results of different GAN models by training a lesion classifier on synthetic data only. In another work, [3] generate skin lesions from ISIC 2018 by translating lesion segmentation masks to images. The resulting images are thus directly associated with ground truth segmentations, which can be leveraged for further applications.

In this paper we present our results for two different types of skin lesions: eczema and moles. For eczema we use a private data set (due to identifying patient information) but for moles we use an established public data set for reproducibility and as an example of the generality of our approach.

Besides technical applications such as data augmentation or the creation of paired data, image transformation also enables domain-specific use cases such as prediction of a skin lesion evolution or the evaluation of aesthetic effects of treatment. With this in mind, we train our GAN models to add or remove eczema from skin pictures pursuing two different strategies: a supervised approach where we use ground truth lesion segmentation masks to target modifications to precisely defined areas as well as an unsupervised process entirely freed from the availability of training data.

2 Materials and Methods

2.1 Data Sets

We conduct experiments on 3 different types of dermatologic images:

Sets of Hands. The first set of experiments is conducted on photos of hands. Each of the 246 individual pairs of hands was photographed from the front and the back side, for a total of 492 photos. They were taken under uniform condition with green background and downscaled to 640 × 480 pixels.

Patches of Skin. Most of the remaining experiments leverage high-resolution photos (3456 × 2304 pixels) of the back side of hands from the EUSZ2 data set collected in the SkinApp project [17]. There are 79 photos available for training and we use a test set of 52 photos to analyze the overfitting of the discriminator. The photos are annotated with segmentations marking the contour of the hands and eczema lesions. From these photos, we extract patches of skin fulfilling the following criteria: a patch consists of skin only (no background) with a specified amount of skin being afflicted with eczema. We create a data set with *healthy skin* patches and a data set with *skin with eczema* patches, where 10–80% of the skin pixels are annotated as eczema. For these experiments, patches of 128×128 pixels are used. This procedure yields 51023 patches of healthy skin and 2872 patches of skin with eczema. Larger patch sizes yield smaller data sets and significantly increase overfitting, especially in the case of skin with eczema.

Skin Lesions. The final data sets consist of dermoscopic images of skin lesions from the ISIC archive 2018 [5,22]. In particular, we generate new lesion images of *Dermatofibroma* (DF) and *Melanoma* (MEL) with 115 and 1113 samples available for training, respectively. These different data set sizes allow to analyze the effects on GAN performance. The original images have varying sizes and are resized to a common resolution of 256×256 pixels.

2.2 Model Architecture

This section describes the architecture of the generator and discriminator models for the experiments. Our models are based on the architecture of DCGAN [19] with the changes described in the following paragraphs. All models are optimized using Adam [16] with a learning rate of $5 \cdot 10^{-5}$ and default moment decays $\beta_1 = 0.9$, $\beta_2 = 0.999$ (values determined experimentally for model convergence). The training was organized in batches of varying size depending on the image resolution and was stopped when the training metrics converged.

Unconditional Generator. The generator for unconditional image synthesis receives a 100-dimensional input vector (drawn independently from a standard Gaussian), which is first passed through a dense layer to produce 64 initial feature maps. The layer's output is reshaped based on the desired aspect ratio of the generated images with lower resolution. Then, a sequence of fractionally-strided convolutions (deconvolutions) increases the image size until the desired output resolution is achieved.

Following common practice, the number of feature maps per convolution are halved at each resolution stage. After each convolution, the output is passed through batch normalization [13] and activated with LeakyReLU [18]. Finally, a regular convolution with 3 output feature maps is activated with tanh to produce the RGB-channels of the generated image.

The hand images generator benefits from unstrided convolutions after each deconvolution to refine the intermediate representations. This is attributed to the comparatively large complexity of these images and does not help with the generation of patches of skin and skin lesions. The size of the initial dense layer and the number of deconvolutions determine the image resolution. Table 1 summarizes the model parametrizations.

Table 1. Unconditional generator: image resolution overview.

Experiment	Dense layers	Deconvolution	Resolution
Full hands (Sect. 3.1)	$20 \times 15 \times 64$	5	640×480
Skin patches (Sect. 3.1)	$8 \times 8 \times 64$	4	128×128
Skin lesions (Sect. 3.1)	$8 \times 8 \times 64$	5	256×256

Image Translation Generator. The image translation model is based on the U-Net architecture [20]: an encoder with increasing number of features, which reduces the image resolution, and a decoder to reverse the process. Additionally, the encoded representation is translated with a sequence of residual blocks [10]. We find experimentally (with the FID score and a qualitative review of the results) that 2 strided convolutions in the encoder and 2 deconvolutions in the decoder yield the best results. Consequently, the residual blocks translate features with a resolution of 32 × 32 pixels. We find that 4 residual blocks are ideal, which is surprisingly low but can be attributed to the fact that the skin images are small and relatively simple. Skip connections between the encoder and the corresponding decoder stages are used as suggested by [14]. These connections forward intermediate features from the encoder that are combined with the decoder features by concatenation.

Finally, we task the image translation generator with *image modification.* To that end, the input image is added to the 3 output channels of the generator, so that it is essentially tasked with generating an image *residual.* The generated residual contains the information to modify the input photo in the desired way.

Discriminator. All experiments leverage the same *multi-scale discriminator* architecture [23]: two individual discriminators process an input image and a downscaled version of the image. Afterwards, their outputs are averaged. This improves the sensibility to low-level details and high-level structures. We observed that more than two discriminators do not improve results, which can be explained by our images' lower resolution when compared with [23].

Both discriminators have the same architecture: a sequence of strided convolutions with batch normalization and LeakyReLU activation, followed by a dense layer with one output neuron to produce the prediction. The features are doubled after each convolution and the number of convolution layers matches the deconvolution layers of the corresponding generators, as summarized in Table 1. All the image translation experiments operate on patches of skin image with 4-convolution discriminators. As the generators produce normalized images, the channels of the real images are also normalized before discrimination.

Model Balance and Selection. The balance between the generator and discriminator is difficult to maintain, as neither should overpower the other [25]. Model balance is adjusted by selecting the number of *initial features* of the generator and discriminator. Table 2 summarizes the initial features of all models in this work's experiments. The ideal numbers of features are determined empirically with the restriction of the available GPU memory.

Besides visual inspection, we minimize the Fréchet Inception Distance (FID) [11] to select the best model. The FID measures the dissimilarity between real and generated images, it is commonly used to quantitatively compare the results of GAN models. In our experiments, this metric works well with unconditional generation, but not with image translations showing that the generator's secondary objective of retaining certain image regions penalizes image realism. Furthermore, we observe that FID scores computed on different data sets should not be compared as the data set's inherent statistics and variability greatly influence the FID scores.

Model selection is additionally guided by the discriminator's predictions confidence and consistency, which indicate whether the discriminator requires additional capacity to adequately distinguish real and generated samples, and thus, to better guide generator learning.

3 Experiments

3.1 Unconditional Dermatology Data Synthesis

The first experiments concern the unconditional generation of dermatology data. The objective is to explore the quality of generated images for different target data sets. The findings indicate the expected performance when the GAN task is not restricted and serves as a baseline for later comparisons with the results of restricted tasks.

Table 2. Initial features for the generator and discriminator models.

Experiment	Generator	Discriminator
Full hands (Sect. 3.1)	512	32
Healthy patches (Sect. 3.1)	1024	128
Eczema patches (Sect. 3.1)	1024	256
Skin lesions (Sect. 3.1)	512	64
Targeted eczema (Sect. 3.2)	1024	256
Untargeted eczema (Sect. 3.3)	1024	256

Sets of Hands. There are two central aspects to the quality of the generated images: high-level structures like anatomy and low-level details like textures. Here, the multi-scale discriminator architecture proves useful, as the two discriminators each focus on one of these aspects. However, many of the generated images still contain visible defects such as hands with more than 5 fingers. These issues are linked to unlikely generator input vectors and can be mitigated using the *truncation trick* [23] to improve the quality of the generated images.

The truncation technique includes the truncation of the input below some a priori defined threshold. Every exceeding component of the input vector is re-sampled. Truncation trades sample variability for quality: aggressive truncation significantly reduces variability, while sample quality increases. We determine empirically that a threshold of 0.1 is suitable for the generation of hands, based on the generated samples and FID scores. These scores are summarized in Table 3. Figure 1 shows the results with a truncation threshold of 0.1.

While the samples do not show great variability, their quality is generally high. The hands' textures look realistic, the side (front or back) of most pairs of hands can be determined in most samples and most hands consist of four fingers and a thumb.

Table 3. Truncation threshold selection with FID score.

Threshold	0.01	0.02	0.05	**0.1**	0.2	0.5	1	None
FID	111.4	94.5	75.0	**69.5**	69.5	70.3	74.1	74.2

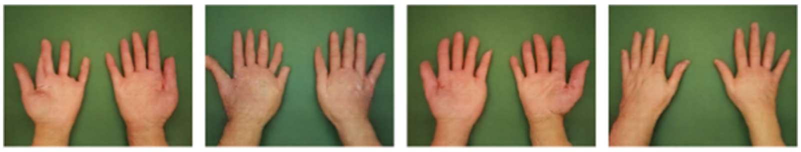


Fig. 1. Samples of the unconditional generation of hands.

This application shows that high-resolution dermatology images can be generated with a relatively small data set. These images could be mistaken for real photos at short glance. The model obtains a FID score of 74.2 without truncation, a significantly lower value than in all other experiments. This indicates that FID scores on different data sets should not be compared.

Patches of Skin. We further experiment with the unconditional generation of images of healthy skin and of skin that contains eczema. These experiments are a prerequisite for later eczema modification experiments.

Healthy Skin. With the large data set of 51023 patches of skin that do not contain any eczema, our GAN is able to generate high-quality images. Samples are shown in Fig. 2. The generated samples look very realistic and are also very diverse. Different types of skin, as well as creases and wrinkles are generated. The selected model achieves a FID score of 538.7.

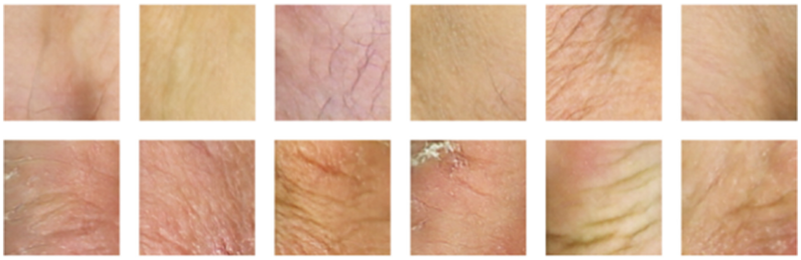


Fig. 2. Samples of the unconditional generation of healthy skin (first line) and skin with eczema (second line).

Skin with Eczema. We observe that the discriminator's task becomes more difficult when classifying patches of skin with eczema, so that the best results are achieved when the discriminator contains more feature maps. Sample results are shown in Fig. 2. The quality of the generated images is comparable with the synthetic healthy skin. The skin is detailed and contains different kinds of wrinkles and eczema. Overall, there are more creases than in the patches of healthy skin, which is attributed to the increased prevalence of eczema in such areas of the hand. The model achieves a FID score of 599.6 for this task.

Perceptual Study. We further evaluate the generated images quantitatively in a perceptual study. The results are presented in Sect. 3.1 along with the analysis of synthetic skin lesion images.

Overfitting. Finally, we analyze the models' overfitting, quantitatively for the discriminator and qualitatively for the generator. For patches of skin with eczema, the discriminator increasingly overfits over the course of the training. Samples from the training set are predicted as real with high likelihood, while testing samples are increasingly being rejected as generated. We observe that this is not the case for the discriminator of healthy skin. As the discriminator for skin with eczema has greater capacity, it is more prone to overfitting. However, we find that overfitting is mainly linked to the data set size. Low-capacity discriminators also overfit to the set of 2872 images, while high-capacity discriminator do not overfit on larger data sets.

We further investigate how the overfitting of the discriminator for patches of skin with eczema impacts the generator. We perform a qualitative assessment of the generator overfitting with the common method of comparing generated samples with their nearest training samples [4,6,15]. In our experiments, the *structural similarity index* [24] yields more similar samples than the *mean squared error*. We find that the generated samples do not contain memorized parts of the training set, so we can conclude that the discriminator's overfitting is not leading the generator to overfit as well.

Skin Lesions. Finally, we generate images of skin lesions. Samples of generated DF and MEL lesions are shown in Fig. 3.

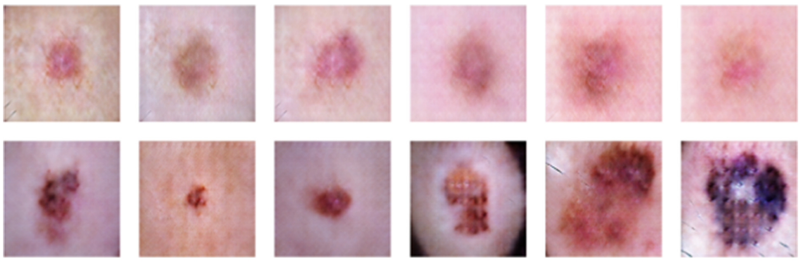


Fig. 3. Samples of the unconditional generation of DF (first line) and MEL (second line) lesions.

Dermatofibroma. While these images resemble the samples of the training set, they lack variability. Furthermore, they show clear tiling artifacts, i.e. patterns that are repeated within a generated image. In this case, the discriminator is trained with only 115 real samples and overfits severely. This visibly impacts the generator: we observe structures, such as lesion shapes or the hairs in the bottom left corners across different samples. With these negative aspects, the generator achieves a FID score of 822.9.

Melanoma. The generated images of MEL lesions contain far greater variability but also suffer from significant tiling. In this case, the generator's FID is 607.8. There is significantly less overfitting, as this data set contains 1113 samples. However, some of the hairs are still repeated. We hypothesize that such specific and distinctive hairs are prone to be copied, as they are rare among the real samples.

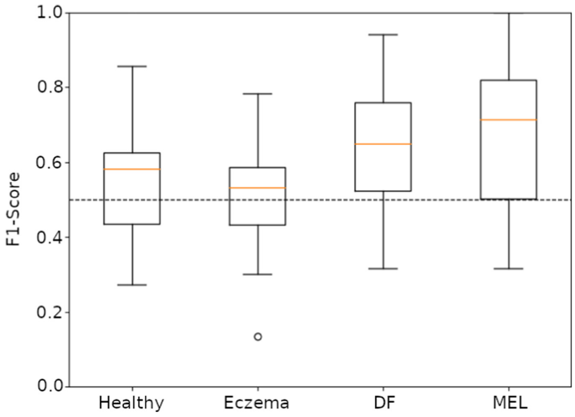


Fig. 4. Perceptual study: the box plots show the three quartiles of the obtained F1-scores for each data set.

Perceptual Study. We assess the realism of the generated patches of skin lesions with a perceptual study, where we ask 104 participants (laymen without prior training) to determine whether a given image is real or generated. The participants are asked to discriminate 20 images from one of four sets: *patches of healthy skin*, *patches of skin with eczema*, *DF lesions*, and *MEL lesions*. They have 2–3 seconds observation time per image and do not receive intermediate feedback. Such experiments are often conducted to assess if the generated images are easily identified [14,21,23]. The classifications are evaluated with the F1-score and the distribution of the results are visualized per data set in Fig. 4. The majority of participants are unable to distinguish real and generated patches of skin, regardless of the presence of eczema: the mean F1-scores are just above random guessing, with 0.58 and 0.53. The third quartiles are also very low, with 0.63 and 0.59. This result confirms that the models are able to generate realistic skin patches. On the other hand, skin lesions are simpler to distinguish, with a mean F1-scores of 0.65 and 0.71. This reflects the observations of the qualitative analysis, where generated lesions look less realistic than synthetic patches of skin. Interestingly, DF lesions are perceived as slightly more realistic than MEL lesions.

3.2 Targeted Eczema Modification

We formulate eczema addition and removal as an image translation task: the generator receives a skin photo and an eczema segmentation mask as input and should either remove or add eczema within the indicated areas. This is performed by generating a residual, which is added to the input image. To encourage pairing between the generator's input and output, its adversarial objective is combined with the *relevancy loss* [1].

The translations are performed between the data sets of skin with and without eczema, two data sets with very different sample sizes. Thus, the set of patches of healthy skin is truncated to 2872 samples, to match the smaller data set. We use additional healthy skin images to train the discriminator for eczema removal, which effectively prevents overfitting. Furthermore, we use the same segmentation with multiple photos of healthy skin. This also helps with generalization, though the effects of this technique are less pronounced.

Eczema Removal. In Fig. 5 we show the translation results of removing eczema from afflicted skin. Columns 3 and 6 still show the same parts of hands as the input photos in columns 1 and 4, but they no longer contain the structures and skin disruptions associated with eczema. However, the generated patches generally lose some fine details such as creases, which are often less visible, compared to the inputs. We observe that the FID score applies poorly to the results of image translation. For these experiments, the FID is often oscillating, in this case between 600 and 1100. Thus, we rely on the visual qualitative evaluation of the generated samples.

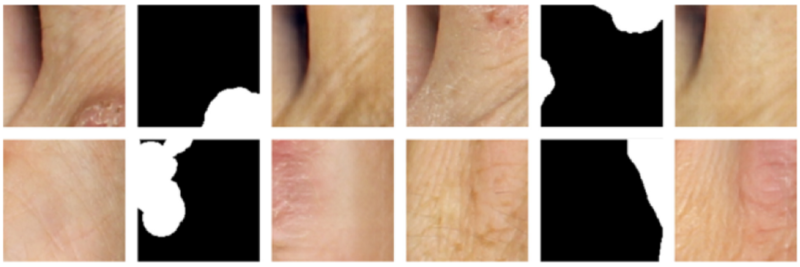


Fig. 5. Eczema removal (first line) and addition (second line) from afflicted skin: columns 1 and 4 show the input photos, columns 2 and 5 the input segmentations and columns 3 and 6 the generation results.

Eczema Addition. We modify photos of healthy skin by adding eczema to specified areas. Figure 5 shows sample results of this translation. The generator again produces realistic images, as we show in columns 3 and 6. Generally, the structures of the skin are retained and fewer details are lost, compared to eczema removal. Further, realistic-looking eczema is placed in the desired parts of the images. These results show that convincing eczema can be in-painted accurately in the indicated locations, which enables applications such as simulating the progression of untreated eczema.

3.3 Untargeted Eczema Modification

We experiment the cyclic translation between patches of skin with and without eczema. No segmentation masks are used and the translations are learned with the completely unsupervised CycleGAN framework [26]. The pairing between generator input and output is achieved with the *cycle consistency loss* [26], which penalizes differences between a generator's input and its reconstruction. While placing a greater emphasis on cycle consistency does increase the pairing, this benefit comes at the cost of reduced sample quality. Sample results of unsupervised eczema modification are shown in Fig. 6.

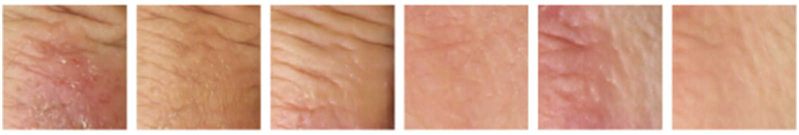


Fig. 6. Unsupervised cyclic eczema transformation: columns 1 and 4 show the sick and healthy input photos, columns 2 and 5 the generated translations without and with eczema and columns 3 and 6 the input reconstructions.

The results are realistic and the original inputs are reasonably reconstructed although some details are missing. This is to be expected, as the generated patches of healthy skin in column 2 should not contain any hints on where or how to in-paint specific eczema. Eczema addition produces realistic-looking lesions, however, it is no longer targeted and can not always be clearly determined.

The loss of details observed in previous translation experiments is barely noticeable here, likely a positive effect of the cycle consistency objective. The metrics of these cyclic translation experiments are more stable than those of the individual translations. For completeness, we mention that the synthetic patches of healthy skin have a FID of 654.7 to the real data, while the synthetic patches of skin with eczema have a FID of 690.2. These scores are reasonably similar to the scores of unconditional generation, with 538.7 and 599.6, respectively.

4 Conclusion

We present different applications of GANs on dermatologic images. First, unconditional image generation is performed successfully with photos of hands and patches of skin in particular. This is also shown for skin patches in the perceptual study. The validity of our approach is therefore confirmed and our initial objective to create realistic synthetic data achieved.

In the case of generated skin lesions, the results do not look as realistic. This could be corrected by further filtering of the images with rare features (such as hair in our particular case), when compared to the other images in the data set. Our analysis shows that the discriminator already overfits with data sets of several thousand images. On the other hand, we only notice overfitting in the generator when using smaller data sets of merely hundreds of samples. Thus, we conclude that the discriminator complexity should be especially controlled when working with small data sets

In the second part of this work, we explore the task of image modification, with eczema addition or removal within a specified area. The obtained results are again visually appealing but we observe that the FID score may be unsuitable to assess the quality of image translation experiments. In particular, we demonstrate the precise addition of eczema to the areas indicated by the segmentation mask. These results open the door for new applications in dermatology such as anomaly detection in a disease appearance or the visualization of the long term aesthetic effects of a disease.

Finally, we also perform domain translation between healthy skin and skin with eczema lesions in an entirely unsupervised experiment. In particular, the eczema removal results may be interesting for future applications, such as weakly-supervised eczema segmentation similar to [1]. This is certainly the most probable case that researchers will encounter as labeling is a costly step. In practice, before labeling is even considered, it is often necessary to first get prototyping results which could be achieved following this approach.

References

1. Andermatt, S., Horváth, A., Pezold, S., Cattin, P.C.: Pathology segmentation using distributional differences to images of healthy origin. CoRR abs/1805.10344 (2018). http://arxiv.org/abs/1805.10344
2. Baur, C., Albarqouni, S., Navab, N.: Melanogans: high resolution skin lesion synthesis with GANs. CoRR abs/1804.04338 (2018). http://arxiv.org/abs/1804.04338
3. Bissoto, A., Perez, F., Valle, E., Avila, S.: Skin lesion synthesis with generative adversarial networks. CoRR abs/1902.03253 (2019). http://arxiv.org/abs/1902.03253
4. Brock, A., Donahue, J., Simonyan, K.: Large scale GAN training for high fidelity natural image synthesis. CoRR abs/1809.11096 (2018). http://arxiv.org/abs/1809.11096
5. Codella, N.C., 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)
6. Denton, E.L., Chintala, S., Szlam, A., Fergus, R.: Deep generative image models using a Laplacian pyramid of adversarial networks. In: Cortes, C., Lawrence, N.D., Lee, D.D., Sugiyama, M., Garnett, R. (eds.) Advances in Neural Information Processing Systems 28, pp. 1486–1494. Curran Associates, Inc. (2015). http://papers.nips.cc/paper/5773-deep-generative-image-models-using-a-laplacian-pyramid-of-adversarial-networks.pdf

7. Frid-Adar, M., Diamant, I., Klang, E., Amitai, M., Goldberger, J., Greenspan, H.: GAN-based synthetic medical image augmentation for increased CNN performance in liver lesion classification. CoRR abs/1803.01229 (2018). http://arxiv.org/abs/1803.01229
8. Goodfellow, I., et al.: Generative adversarial nets. In: Ghahramani, Z., Welling, M., Cortes, C., Lawrence, N.D., Weinberger, K.Q. (eds.) Advances in Neural Information Processing Systems 27, pp. 2672–2680. Curran Associates, Inc. (2014). http://papers.nips.cc/paper/5423-generative-adversarial-nets.pdf
9. Guibas, J.T., Virdi, T.S., Li, P.S.: Synthetic medical images from dual generative adversarial networks. CoRR abs/1709.01872 (2017). http://arxiv.org/abs/1709.01872
10. He, K., Zhang, X., Ren, S., Sun, J.: Deep residual learning for image recognition. CoRR abs/1512.03385 (2015). http://arxiv.org/abs/1512.03385
11. Heusel, M., Ramsauer, H., Unterthiner, T., Nessler, B., Klambauer, G., Hochreiter, S.: GANs trained by a two time-scale update rule converge to a nash equilibrium. CoRR abs/1706.08500 (2017). http://arxiv.org/abs/1706.08500
12. Hiasa, Y., et al.: Cross-modality image synthesis from unpaired data using cyclegan: effects of gradient consistency loss and training data size. CoRR abs/1803.06629 (2018). http://arxiv.org/abs/1803.06629
13. Ioffe, S., Szegedy, C.: Batch normalization: accelerating deep network training by reducing internal covariate shift. CoRR abs/1502.03167 (2015). http://arxiv.org/abs/1502.03167
14. Isola, P., Zhu, J.Y., Zhou, T., Efros, A.A.: Image-to-image translation with conditional adversarial networks. In: The IEEE Conference on Computer Vision and Pattern Recognition (CVPR), pp. 1125–1134 (2017)
15. Karras, T., Aila, T., Laine, S., Lehtinen, J.: Progressive growing of GANs for improved quality, stability, and variation. CoRR abs/1710.10196 (2017). http://arxiv.org/abs/1710.10196
16. Kingma, D.P., Ba, J.: Adam: a method for stochastic optimization. CoRR abs/1412.6980 (2014). http://arxiv.org/abs/1412.6980
17. Koller, T., Schnürle, S., vor der Brück, T., Christen, R., Pouly, M.: Skinapp deeplearning. Technical report, Lucerne University of Applied Sciences, September 2018
18. Maas, A.L., Hannun, A.Y., Ng, A.Y.: Rectifier nonlinearities improve neural network acoustic models. In: ICML Workshop on Deep Learning for Audio, Speech and Language Processing, p. 3 (2013)
19. Radford, A., Metz, L., Chintala, S.: Unsupervised representation learning with deep convolutional generative adversarial networks. CoRR abs/1511.06434 (2015). http://arxiv.org/abs/1511.06434
20. Ronneberger, O., Fischer, P., Brox, T.: U-net: Convolutional networks for biomedical image segmentation. CoRR abs/1505.04597 (2015). http://arxiv.org/abs/1505.04597
21. Salimans, T., Goodfellow, I.J., Zaremba, W., Cheung, V., Radford, A., Chen, X.: Improved techniques for training GANs. CoRR abs/1606.03498 (2016). http://arxiv.org/abs/1606.03498
22. Tschandl, P., Rosendahl, C., Kittler, H.: The ham10000 dataset, a large collection of multi-source dermatoscopic images of common pigmented skin lesions. Sci. Data **5**, 180161 (2018)
23. Wang, T., Liu, M., Zhu, J., Tao, A., Kautz, J., Catanzaro, B.: High-resolution image synthesis and semantic manipulation with conditional GANs. CoRR abs/1711.11585 (2017). http://arxiv.org/abs/1711.11585

24. Wang, Z., Bovik, A.C., Sheikh, H.R., Simoncelli, E.P.: Image quality assessment: from error visibility to structural similarity. IEEE Trans. Image Process. **13**(4), 600–612 (2004). https://doi.org/10.1109/TIP.2003.819861
25. Yi, X., Walia, E., Babyn, P.: Generative adversarial network in medical imaging: a review. ArXiv e-prints (2018)
26. Zhu, J., Park, T., Isola, P., Efros, A.A.: Unpaired image-to-image translation using cycle-consistent adversarial networks. CoRR abs/1703.10593 (2017). http://arxiv.org/abs/1703.10593

Chapter 9

Teledermatology

9.1 Introduction

The diversity and prevalence of skin diseases [50] naturally induce a large demand of dermatology services that has been further fueled by public campaigns raising awareness towards certain skin conditions such as skin cancer [185]. However, this demand is currently not satisfactorily fulfilled given the general shortage of dermatologists [97, 161] and the inherent artisanal nature of dermatology services, which cannot be effectively scaled. Dermatologist education is both demanding and expensive and with an increasing proportion of the world population gaining access to health services, the lack of experts is felt stronger in every country.

Thanks to the visual nature of skin diseases, patients' condition can easily be documented with imaging and questionnaires, making teledermatology a promising approach, with first reports in the scientific literature dating back to 1995 [153]. Since then, several reviews have extensively analyzed teledermatology publications [102, 149, 191]. As the sub-field of telemedicine focusing on skin conditions, teledermatology aims to provide remote dermatology services such as diagnosis [208], triage [35] or follow-up [36] to patients by leveraging available means of communication. Its adoption, both by experts and patients, has been propelled by the COVID pandemic [51] and the general availability of mobile phones able to capture high-quality photographs.

There are several categories of teledermatology, which are distinguished based on the information flow and the actors involved [191]. Situations involving only patients and nurses or general practitioners are called primary teledermatology. When the patient's information needs to be forwarded to a dermatologist, this becomes secondary teledermatology. In challenging cases, more specialized experts (e.g. dermatopathologist) may be contacted, in which case we talk about tertiary teledermatology. Patient-assisted teledermatology occurs when the patient is in direct interaction with either actors. Another defining characteristics of teledermatology is how information is exchanged between the involved actors. Three main modalities were identified: asynchronous store-and-forward (the most frequent [21, 191]), real-time and hybrid. While real-time teledermatology consultations offer direct interactions and enable dermatologists to better analyze patients' condition, they are more time-consuming and require more technical resources. In comparison, store-and-forward teledermatology adjusts better with experts' schedule, is cheaper for patients and more scalable. Exchanged images are usually of better resolution than what can be achieved with real-time consultations. Hybrid teledermatology is a combination of both modalities adapted to fit health centres' workflows and specific business models.

Teledermatology benefits all stakeholders of the dermatology sector. It enables patients in remote areas or with limited mobility to receive care directly at home. With store-and-forward teledermatology, dermatologists are more flexible and can optimize their schedule. Triage shortens waiting lists for patients and avoid unnecessary face-to-face consultations, reducing healthcare costs as a whole. Furthermore, teledermatology assists in the continuous education of primary care health workers and even dermatologists, who are confronted with additional patients and can seek assistance of more experienced peers in challenging cases. Overall, the result is a general, cost-effective scale up of dermatology services [200] (even without the use of artificial intelligence).

Teledermatology also face some challenges and drawbacks. On the technical level, it is dependent on the quality of the means of communication: a poor internet connection will decrease user experience for both patients and dermatologists. In comparison to face-to-face consultations, dermatologists do not have access to the full patient condition. The available information is limited and may introduce bias, causing dermatologists to overlook important considerations. In skin cancer screening, for example, patients may not include moles because they simply ignore their existence. Furthermore, some information cannot be communicated, such as the one acquired by palpation. An inherent downside of teledermatology is the reduction of interactions between patients and dermatologists. This results in less empathy and more awkward exchanges, as patients cannot ask questions naturally. Finally, sharing information remotely introduces new challenges regarding security and privacy of patient data.

Teledermatology combined with artificial intelligence (AI), has the potential to further scale up dermatology health services [149]. However, the actual use of deep learning models in teledermatology settings is not yet common due to several challenges (cf. section 4.3) and remains open research with ongoing initiatives [85]. The principal exception is skin cancer screening applications [42, 57, 172]. In this context, dermoscopy images were shown to be beneficial [204] and can be acquired using commercial teledermoscopy mobile phone accessories. A study confirmed that dermoscopy deep learning models (DLMs) achieved similar performance on images acquired in teledermatology or in clinical settings [203]. However, none of the public skin cancer screening platforms offer legal guarantees on the provided feedback, which is never framed as a real diagnosis. So far, other applications consist in training deep learning models on teledermatology patients' data [87, 113, 138]. Again, none of them are deployed in practice except for secondary non-medical applications such as the organization of dermatology databases [207].

In the next sections, we present a research initiative aiming to create an AI-assisted teledermatology platform in Africa.

9.2 The PASSION Project: Pediatrics in Africa

The burden of skin diseases is especially heavy in Africa, where 21% to 87% of children suffer from skin diseases [146] and the shortage of dermatologists is very severe [171]. Common skin conditions are often left untreated or worse, mistreated following wrong diagnoses, which leads to critical degradation of patients' quality of life and adverse consequences physically, psychologically and socially. In a joint initiative funded by the Fondation Botnar[1], the University Hospital Basel together with the dermatology centers of Tanzania, Madagascar and Guinea, launched in 2020 the PASSION project (Pediatrics in Africa — Enabling Wireless Diagnosis for Common Skin Diseases) [85], which aims to democratize access to dermatology services in Africa with a particular focus on children patients. The overarching goal of the project is to create an AI-assisted triage and diagnosis teledermatology platform, which would empower general practitioners and nurses to diagnose and treat skin conditions. Six conditions have been selected among the most prevalent [93, 99]: atopic dermatitis (eczema), mycosis, bacterial infections, impetigo, insect bites, scabies, and other/control. Our platform is envisioned to serve both as an accessible primary point of care and to support local healthcare service providers. Via a mobile phone application, patients will submit their pictures, reply to questionnaires and receive the most likely diagnosis, counseling and follow-up recommendations.

The project is composed of three successive phases, described in figure 9.1. First, the data necessary to train DLMs is acquired via standard consultations performed in partner clinical centers and teledermatology consultations performed by primary care workers in remote areas. Due to the COVID pandemic and local challenges described in the next section, this phase is still ongoing. We have developed a prototype DLM mainly for eczema. Once the laboratory performance is sufficient for deployment, the second phase will consist in running standard store-and-forward teledermatology consultations and compare the DLM predictions with experts' decisions. The success of these consultations will be evaluated after a three months follow-up period. In the final phase, provided phase two's DLMs performance was on par with experts, we will scale up the platform deployment and perform automatic triage of patients with treatment recommendation. Success of consultations will be confirmed past a twelve months follow-up period.

The project has been designed in a modular organization to maximize its possible outcomes. The first module is the creation of a database of skin conditions from Africa, which will be made available to researchers after anonymization. The second is to provide a store-and-forward teledermatology platform, which is currently tested in our partner healthcare centers. The third is to create DLMs that can perform triage of skin conditions on patients with type IV to VI on the Fitzpatrick scale. We expect that the PASSION project will contribute to the general improvement of the local health situation by improving the diagnosis and treatment of common skin conditions. Upon the project completion, the models and teledermatology platform will be integrated in

[1] https://www.fondationbotnar.org/ (Accessed: 2nd February 2023)

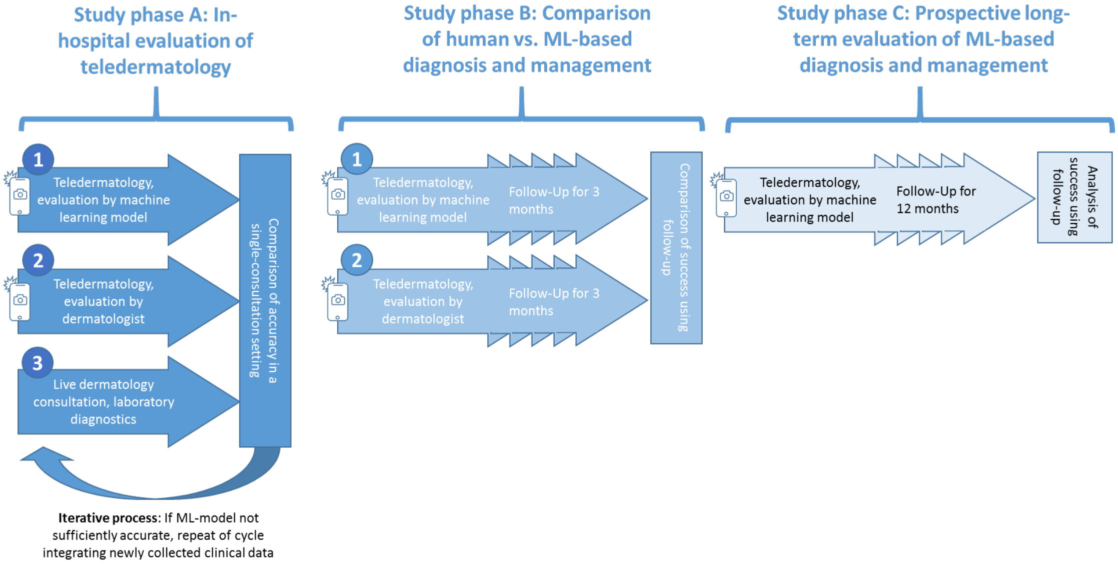

Study phase A: In-hospital evaluation of teledermatology
Study phase B: Comparison of human vs. ML-based diagnosis and management
Study phase C: Prospective long-term evaluation of ML-based diagnosis and management
1
Teledermatology, evaluation by machine learning model
2
Teledermatology, evaluation by dermatologist
3
Live dermatology consultation, laboratory diagnostics
Comparison of accuracy in a single-consultation setting
Iterative process: If ML-model not sufficiently accurate, repeat of cycle integrating newly collected clinical data
1
Teledermatology, evaluation by machine learning model
Follow-Up for 3 months
2
Teledermatology, evaluation by dermatologist
Follow-Up for 3 months
Comparison of success using follow-up
Teledermatology, evaluation by machine learning model
Follow-Up for 12 months
Analysis of success using follow-up

Figure 9.1: Description of the PASSION [85] project's phases.

established healthcare provider platforms.

9.3 Medical Image Collection in Sub-Saharan Africa

This short conference paper was submitted but rejected at the 2022 Swiss Conference on Data Science. Peer reviews were positive but required additional experiments with DLMs, which could not be performed at the time due to the lack of data. This issue should be solved in the near future, enabling the resubmission of the paper. In this work, we discuss the challenges faced in the data collection phase of our teledermatology research initiative and present the results achieved by our prototype DLMs for eczema triage.

Medical Image Collection in Sub-Saharan Africa

Ludovic Amruthalingam[§]
University of Basel
Basel, Switzerland
ludovic.amruthalingam@unibas.ch

Philippe Gottfrois[§]
University of Basel
Basel, Switzerland
philippe.gottfrois@usb.ch

Alvaro Gonzalez-Jimenez
University of Basel
Basel, Switzerland
alvaro.gonzalezjimenez@unibas.ch

Rapelanoro R. Fahafahantsoa
Madagascar Society of Dermatology
Antananarivo, Madagascar
frapelanoro@gmail.com

Ibrahima Traoré
Dermatology Clinic
Konakry, Guinea
trachimi@yahoo.fr

Marc Pouly
Lucerne University of Applied Sciences and Arts
Rotkreuz, Switzerland
marc.pouly@hslu.ch

Alexander A. Navarini
University Hospital of Basel
Basel, Switzerland
alexander.navarini@usb.ch

***Abstract*—Skin diseases in Africa affect up to 87% of children. Good treatments are available for the most prevalent diseases, provided they are diagnosed at an early stage. However, there is a severe shortage of dermatologists and lack of the necessary health services. In 2020, a joint field initiative in Tanzania, Madagascar and Guinea was launched to democratize access to dermatology using AI and telemedicine to perform patient triage and diagnosis on the 4 most prevalent skin conditions. We report challenges and their successful mitigation on federated data collection in developing countries and under exceptional conditions of an ongoing global pandemic.**

***Index Terms*—health, dermatology, Africa, telemedicine, artificial intelligence**

I. Introduction

In Africa, up to 87% of children suffer from a skin disease [1] but the local healthcare systems cannot provide the necessary experts to diagnose and treat them. While in Switzerland, there is a ratio higher than one dermatologist for 20000 people [2], this ratio falls to less than 1 dermatologist per million people in several African countries [3]. Consequently, common skin conditions are often misdiagnosed or mistreated leading to severe impairment of quality of life and chronic morbidity.

Atopic dermatitis/eczema, bacterial and fungal skin infections, scabies account for more than 80% of pediatric skin disease patients in Tanzania Standard treatments are available for each disease and often prevent further complications, if initiated early enough [4], [5].

The PASSION project[1], launched in 2020, is a joint initiative by the University Hospital Basel and several dermatology centers in Tanzania, Madagascar and Guinea funded by the Fondation Botnar[2]. It aims to create an AI-assisted teledermatology platform for pediatric skin conditions in sub-Saharan countries. Triage of common and easily treatable skin conditions shall be performed (semi-)automatically to allow dermatologists to focus on the most severe and complex cases.

II. Data Collection in Sub-Saharan Africa

Training a machine learning based diagnostic system requires a sufficient amount of high-quality data, which, in this project, proved particularly hard to acquire. In recent years, several public benchmark datasets of skin lesions were made available: the most established is ISIC and consists of dermatoscopic images, the most recent and largest is [6]. However, these and all other publicly available datasets differ significantly from the PASSION use case:

- Skin pigmentation: public datasets were mainly collected in European and Asian countries. They have very limited pigmentation range, usually Fitzpatrick types 1 to 3, whereas in sub-Saharan countries type 4 and 5 are most frequent.
- Disease state: due to the lack of specialists, skin conditions are diagnosed at rather late stages of their evolution, when the patient's life is already strongly impaired. This is a fundamental difference from typical clinical datasets, which mainly contain cases of patients already under dermatologist care.
- Diagnoses: in contrast to Europe, infectious conditions rank among the top reasons why people consult a medical doctor in Africa. We cannot expect a similar distribution of diagnoses in datasets collected in different parts of the world.

A. Practical Challenges and Measures Taken

An international medical data collection initiative must respect the different regulations of the involved countries. While strict regulation may impede data collection and research use, lax legislation comes with grey areas with room for interpretation and risks of future restrictive evolution. The project team imposed that all involved partners must follow GDPR compliant procedures in order to have compatible processes in all activities. Moreover, this ensured that all participants were treated fairly concerning global standards. Finally, by adopting the strictest regulations, the project is more likely to be future proof against the evolution of legislation.

[§] Equal contribution
[1] https://www.telederm.ai
[2] https://www.fondationbotnar.org

Establishing trust with local specialists and patients is paramount. Global collaboration can be perceived as neo-colonialism with richer countries looting data from emerging countries. Collecting pictures can also be taken as offense by certain populations. Consequently, local specialists can only do successful data collection with established trust relationships and are best aware of their patient's cultural background and tradition.

Transparency and fairness are guiding principles in this initiative. All partners are directly involved in steering and presenting the project to their local specialists and healthcare workers. With the above mentioned ratio of one dermatologist per million people in mind, we cannot afford a larger dropout of local specialists. Data collection must be as effortless as possible and smoothly integrated into their daily routine. Huge efforts are being made to ease and structure the on-boarding process, that specialists can communicate in their preferred language, etc.

Every hospital collects data for its own purposes, but unfortunately this existing data cannot be used. Typical issues concern the lack of patient consent; inconsistent diagnostic, patient and meta-data; incompatibilities related to file formats and database systems; non-standardized images taken under varying conditions and with heterogeneous devices, etc. However, based on this analysis a standardized protocol could be established for image capturing and documentation of diagnoses including recommendations on relevant metadata to be assessed. In particular, specialists of the International Society of Teledermatology provided guidelines and training for picture taking.

Finally, a dedicated app has been created to implement this standardized data collection protocol and further reward local doctors based on their contribution. As internet connection and electricity are often unstable and weather dependent, much attention was put on service and data transfer reliability. Mobile phones are widespread in Africa [7], but devices and operating systems are extremely heterogeneous. App requirements and battery consumption needed to be as low as possible, image resolution and low bandwidth balanced, etc. The app can store images offline until an internet connection and sufficient battery charge is available again.

B. *The COVID-19 Pandemic*

COVID-19 stroked Europe and Asia two months after the start of the project, which drastically handicapped the so important trust building process due to reduced travelling possibilities. Patients afraid of infection started to avoid medical facilities and dermatologists where progressively reassigned to COVID related tasks or became infected themselves. All this impaired the first year of data collection and almost brought the project to a complete full stop.

We bridged this period by acquiring existing private and anonymized datasets to at least start the development of early machine learning models while at the same time setting up and institutionalize data quality processing with the support of dermatology imagery specialist.

C. *Early Results of AI Models*

We developed two ResNet based models, one with data acquired in the PASSION project (704 pictures, model A) and the other with data available from Swiss hospitals (13748 pictures, model B). Due to the as yet unbalanced distribution of diagnoses, we can only predict eczema against other conditions for the moment. Model A achieves an F1 score of 0.87 while model B achieves 0.79 on unseen test data.

III. Conclusion and Outlook

Organizing a project spanning two continents led to multiple challenges and a series of unexpected problems had to be overcome. We foresaw many problems related to technical and cultural aspects, differences between developing countries and central Europe, but still could have allowed for more project time to build up trust, personal relationships and navigate cultural differences. However, a global pandemic with a complete breakdown of health systems, patients refusing consultation, etc. was definitely not on the risk map. However, thanks to a decentralized project organisation and the engagement and enthusiasm of local partners, the project could recover pretty much immediately from these setbacks. We are now accumulating high-quality data of a kind that research has not seen before. Once it reaches a sufficient size this database will be made available to the research community after thorough anonymization. We hope to be able to soon deploy a machine learning based teledermatology service to help children in remote areas of sub-Saharan Africa.

Acknowledgment

We thank our partners for their cooperation and the Fondation Botnar for their trust and support.

References

[1] World Health Organization, "Epidemiology and management of common skin diseases in children in developing countries" in WHO (2005).

[2] S. Hostettlera, E. Kraft, "Statistique médicale 2019 de la FMH: forte dépendance de l'étranger" in Bull Med Suisses (2020)

[3] E. Lauressergues in "Deuxièmes assises africaines de télé-dermatologie", https://sotoderm.org/2019/07/23/deuxiemes-assises-africaines-de-tele-dermatologie (2019)

[4] O.S. Katibi, N.C. Dlova, A.V. Chateau, A. Mosam, "The prevalence of paediatric skin conditions at a dermatology clinic in KwaZulu-Natal Province over a 3-month period", South African Journal of Child Health (2016)

[5] S.K. Kiprono, J.W. Muchunu, J.E. Masenga, "Skin diseases in pediatric patients attending a tertiary dermatology hospital in Northern Tanzania: a cross-sectional study", BMC dermatology (2015). https://doi.org/10.1186/s12895-015-0035-9

[6] Liu, Y., Jain, A., Eng, C. et al. A deep learning system for differential diagnosis of skin diseases. Nat Med 26, 900–908 (2020). https://doi.org/10.1038/s41591-020-0842-3

[7] Tanzania Daily News, "Tanzania: Smartphones Push Up Internet Penetration" https://allafrica.com/stories/201709080283.html (2017)

Chapter 10

Training and Evaluation Framework

Our digital dermatology lab is composed of interdisciplinary researchers with mainly medical and computer science backgrounds. Regularly, our team is reinforced by master students, mainly from the medical faculty, who are eager to participate in the development of deep learning based dermatology applications. While their medical training enables them to curate the training datasets, they lack the engineering skills to develop deep learning models (DLMs) and are completely reliant on other researchers with the necessary technical skills. To reduce the technical entry barrier, we have developed a modular framework based on the PyTorch [150] and Fastai [84] libraries simplifying the development of typical image-based dermatology applications, which usually follow this workflow: dataset preparation and preprocessing, training and evaluation of classification and segmentation DLMs, inference on new data. In this chapter, we briefly present how our framework can assist at each of these steps and conclude with planned extensions.

10.1 Dataset Preparation and Preprocessing

Splitting Once a dataset has been labeled, the first step is to split it into a training and test dataset that will be used to train and evaluate DLMs performance. This can be performed with the `general/split_dataset.py` script, which also ensures that no patient leaks between the test and train sets occurs.

Patching Due either to labeling costs or scarce data, we are often dealing with small datasets for deep learning standards. In certain applications, dividing images into multiple smaller patches can be useful to leverage the full image resolution and increase the effective size of the training dataset. This can be achieved using the `general/PatchExtractor.py` script, which handles multiple fixed square patch sizes or dynamic patch sizes based on a specified image resolution.

Reducing semantic class imbalance In segmentation tasks, we face situations where the imbalance between semantic classes can be too high for effective training even when using adapted losses. The script `segmentation/crop_to_thresh.py` creates a second version of the training set by (randomly) cropping each images around the chosen semantic classes. The script tries to ensure that the ratio between classes matches the specified threshold.

Encryption When the training machine is located outside the hospital premises, all medical data must be encrypted. This can be performed using the `general/crypto.py` script.

10.2 Training and Performance Evaluation

The framework is developed in the object-oriented programming paradigm. The base class, `FastaiTrainer`, fully abstracts the type of data as well as the task and defines the following aspects:

- Common training arguments such as batch size, number of epochs, initial learning rate, etc.
- Procedure to split the training set and apply cross-validation.
- Generic training procedure, which consists in two phases: frozen training of the last layer, then training of the full DLM with one cycle scheduling of the learning rate [184].
- Prediction and correction of weak labels to leverage unlabeled data when it is available.
- Procedure to alternate training between strongly labeled data (labels provided by experts) and weakly labeled data (labels predicted during training).
- Prediction of the test set labels together with a generic procedure to evaluate performance.

Specialized for images, the class `ImageTrainer` inherits from `FastaiTrainer` and defines:

- Image specific arguments such as the input size, image location, specific losses, etc.
- Image loading with generic image dataloader. Lazy loading is performed when images are not encrypted.
- Splitting the training set for cross-validation, ensuring no leaks of patched images between sets.

- Procedure to apply progressive resizing of the training images.
- Generic metrics computations with visualization plots.
- Logging training losses and validation metrics with Tensorboard [1].

Finally, `ImageClassificationTrainer` and `ImageSegmentationTrainer` inherit from `ImageTrainer` and redefine:

- Task-specific arguments such as location of labels, metrics choice, etc.
- Loading expert's labels with customized dataloaders.
- Creation of DLM based on popular architectures.
- Procedure to update predicted weak labels in the course of training.
- Task specific metrics.

To train a DLM, students can readily use the task-specific classes, ignoring implementation details, and evaluate different parameters' configuration with the training logs and performance reports. Depending on the project, it may be needed to redefine aspects such as model creation and data loading within a child class, which can be performed together with a more experienced researcher.

10.3 Inference

When it comes to DLM evaluation and testing, it is also useful to perform inference on additional data and evaluate the predictions' quality manually. Based on a similar object-oriented approach, we created a base class `ImageInference` to perform the following generic operations:

- Data loading.
- DLM loading with trained weights.
- Inference procedure

Inheriting from this base class, we defined two classes `ImageClassification-Inference` and `ImageSegmentationInference`, which specify task-specific arguments and visualization procedures for DLM predictions. Students can perform inference using these classes similarly to how they perform DLM training.

10.4 Planned Extensions

Thanks to its object-oriented implementation, the presented framework can be extended to different settings such as the multimodal inputs used in chapter 6 and even to other types of tasks. In the future, we plan to extend the framework with instance segmentation and object detection.

In this thesis, we focused solely on images and did not leverage text data from medical reports. Dermatologists have always documented patients' condition with textual descriptions, and there is great potential in creating natural language processing DLMs based on this data. Similarly to how the framework was created for images, its base class `FastaiTrainer` could be extended to handle natural language data with task-specific child classes such as amamnesis summary generation.

Chapter 11

Discussion and Conclusion

Discussion

The main objective of this thesis is to automate and improve parts of dermatologists' daily workflow. We present deep learning models leveraging dermatologists' knowledge to make their decision processes more understandable or at least verifiable so that clinical adoption is simplified. We focused on a fundamental aspect of dermatologists' analysis of skin conditions that can be inferred from pictures only: the dermatological description of lesions. In particular, lesion location, morphology, distribution, counts and surface estimation.

Any lesion description starts with its location. We created a method that can generate a map of anatomical regions from patient pictures, a problem not researched so far in the literature. With a precision ranging from the main regions of the human body, to every single anatomical units of the hands or the ears, the method can assist dermatologists in lesion documentation or be combined with other deep learning applications such as lesion detection to produce detailed reports. Furthermore, the anatomy maps can be used to perform targeted image retrieval for specific body regions in large hospital databases.

The differential diagnosis processes followed by dermatologists are based on the features collected in the dermatological description of lesions. We proposed a method to combine lesion location with image features and showed that it could improve the performance of differential diagnosis deep learning models. Similarly, we evaluated the same approach with the morphology of lesions and observed a significant performance improvement. Both approaches were never attempted before. These features are easier to determine for clinicians than the lesion differential diagnosis and illustrate the potential of combining expert's knowledge with the capacity of deep learning models to automatically extract relevant information. The features can be verified and updated with ease, enabling clinicians to interact with the deep learning system and evaluate the impact of their changes, as well as improving their understanding of the models' decision process.

The distribution, counts and surface estimation of lesions can be determined from their segmentation in patients' pictures. In this work, we proposed a method able to segment palmoplantar pustular psoriasis, ichthyosis with confetti and hand eczema lesions, enabling automated quantification of their lesions for a more objective severity assessment. These diseases have different types of distribution showing the generality of our method, which is also applicable in restricted data availability regimes (e.g. context with rare diseases) as we showed in the case of ichthyosis with confetti. Furthermore, we were able to combine the hand anatomy and hand eczema models to produce the first automated analysis of the disease anatomical stratification.

With the increasingly restrictive legal environment surrounding medical data, researchers must find alternative approaches to share research datasets and enable peers to reproduce results and compare different methods. The generation of artificial data from private datasets being a promising solution to this issue, we illustrated how generative adversarial networks could be trained to produce synthetic data for the main imaging modalities used in dermatology, namely photography and dermoscopy.

One of the broader benefits that deep learning applications bring to society is better access to dermatology services. Following this ideal, we launched the first research teledermatology initiative in Sub-Saharan Africa, aiming to support primary care with semi-automatic triage of the six most prevalent local skin conditions. Despite the pandemic, we were able to establish collaboration with several health centers in three different sub-Saharan countries and launch an international data collection effort. This data will be made available after thorough anonymization to researchers and enable the development of useful deep learning applications for local patients.

The projects covered in this thesis fostered the development of a generic deep learning model training and evaluation framework based on the Fastai [84] and PyTorch [150] libraries. The main motivation behind this effort was to enable collaborators with non-technical backgrounds to participate in the model development process, while requiring the least programming skills as possible. Currently, classification and segmentation tasks are supported but the framework is designed to be extendable to new tasks. We plan to release the code for this framework in the near future.

Future Work

We are already working towards automated anatomy mapping of the full human body. Once achieved, it will enable the extension of our anatomical stratification approach to generalized diseases such as psoriasis or vitiligo. Our aim is to establish an anatomical distribution signature of the different diseases and evaluate how it can support their differential diagnosis.

So far, our experiments on the development of differential diagnosis deep learning models were limited to the separate use of lesion location and morphology, mainly due to the lack of annotated data. We are planning to extend this approach to other features from dermatological description of skin lesion, with the aim to reproduce and automate parts of dermatologists' current differential diagnosis processes.

Label scarcity and the associated challenges was encountered in every project. To mitigate this issue, we pretrained models on ImageNet [44], a large object recognition dataset unrelated to the dermatology domain. We are currently testing semi-supervised methods to pretrain our models on large unlabeled dermatology datasets and will determine whether this brings a general improvement. This approach goes in the direction of learning from data directly rather than only from experts' knowledge, which remains subjective and is usually limited to the state of the art in dermatology at best. Another idea we plan to evaluate is the combined use of original and synthetic data to train models with larger datasets.

While the deep learning models presented in this work achieved satisfactory performance, they were evaluated on pictures captured in hospitals and selected for their relative standardization. These models cannot be applied in too different settings without compromising their performance. Recently, we could start working on a different data modality, mainly 3D full-body scans of very high resolution. With these scans, we will be able to arbitrarily generate our datasets while simulating varied capturing conditions. This should result in more robust and generally applicable models.

One of the limitation of this thesis was to consider only image data. Yet, dermatologists have been documenting diseases with text and drawings long before they could photograph them. Thus, a large source of data has remained untapped so far, also in the research literature. We plan to start training deep learning models to analyze patients medical reports with applications such as the automated identification of relevant information in patients' anamnesis.

Finally, all deep learning applications need to be tested and validated in prospective studies, which is one of the requirements to legally establish them as medical device and enable their deployment in practice. We plan to start a prospective study for our hand eczema anatomical stratification approach. This will allow us to correlate the model predictions with clinical severity grading systems and collect the necessary data to fully automate them.

Conclusion

In this work, we tackled two of the most important tasks in diagnostic dermatology: disease differential diagnosis and severity grading. We showed that letting deep learning models leverage experts' knowledge was beneficial for differential diagnosis and proposed a robust method to segment various skin diseases. Our anatomy mapping approach synergizes well with lesion detection and segmentation applications, improving their predictions' clinical relevance and enabling new research analysis. All methods presented in this thesis either produce easily verifiable predictions or in the case of differential diagnosis, offer clinicians a new level of interactivity. With our teledermatology initiative, we started acquiring the necessary data to adapt our prototype models from clinical settings to real-world conditions with the aim to improve field healthcare conditions and benefit society. To conclude, our thesis takes a step towards the future of dermatology practice: AI-assisted dermatologists.

Bibliography

[1] M. Abadi, A. Agarwal, P. Barham, E. Brevdo, Z. Chen, C. Citro, G. S. Corrado, A. Davis, J. Dean, M. Devin, et al. Tensorflow: Large-scale machine learning on heterogeneous distributed systems. *arXiv preprint arXiv:1603.04467*, 2016.

[2] I. S. A. Abdelhalim, M. F. Mohamed, and Y. B. Mahdy. Data augmentation for skin lesion using self-attention based progressive generative adversarial network. *Expert Systems with Applications*, 165:113922, 2021.

[3] A. S. Adamson and A. Smith. Machine learning and health care disparities in dermatology. *JAMA dermatology*, 154(11):1247–1248, 2018.

[4] A. Adelekun, G. Onyekaba, and J. B. Lipoff. Skin color in dermatology textbooks: an updated evaluation and analysis. *Journal of the American Academy of Dermatology*, 84(1):194–196, 2021.

[5] C. S. Ahn, L. Culp, W. W. Huang, S. A. Davis, and S. R. Feldman. Adherence in dermatology. *Journal of Dermatological Treatment*, 28(2):94–103, 2017.

[6] S. F. Aijaz, S. J. Khan, F. Azim, C. S. Shakeel, and U. Hassan. Deep learning application for effective classification of different types of psoriasis. *Journal of Healthcare Engineering*, 2022, 2022.

[7] L. Alzubaidi, M. A. Fadhel, S. R. Oleiwi, O. Al-Shamma, and J. Zhang. Dfu_qutnet: diabetic foot ulcer classification using novel deep convolutional neural network. *Multimedia Tools and Applications*, 79(21):15655–15677, 2020.

[8] L. Alzubaidi, J. Zhang, A. J. Humaidi, A. Al-Dujaili, Y. Duan, O. Al-Shamma, J. Santamaría, M. A. Fadhel, M. Al-Amidie, and L. Farhan. Review of deep learning: Concepts, cnn architectures, challenges, applications, future directions. *Journal of big Data*, 8(1):1–74, 2021.

[9] American Academy of Dermatology. American academy of dermatology position statement on augmented intelligence (aui). `https://server.aad.org/Forms/Policies/Uploads/PS/PS-Augmented%20Intelligence.pdf`, 2019. Accessed: 2nd February 2023.

[10] L. Amruthalingam, O. Buerzle, P. Gottfrois, A. Jimenez Gonzalez, A. Roth, T. Koller, M. Pouly, and A. Navarini A. Quantification of efflorescences in pustular psoriasis using deep learning. *Healthcare Informatics Research*, 28(3):222–230, 2022.

[11] L. Amruthalingam, P. Gottfrois, A. Gonzalez Jimenez, B. Gökduman, M. Kunz, T. Koller, D. Consortium, M. Pouly, and A. Navarini. Improved diagnosis by automated macro-and micro-anatomical region mapping of skin photographs. *Journal of the European Academy of Dermatology and Venereology*, 2022.

[12] L. Amruthalingam, N. Mang, P. Gottfrois, A. Gonzalez Jimenez, J.-T. Maul, M. Kunz, M. Pouly, and A. A. Navarini. Objective hand eczema severity assessment with automated lesion anatomical stratification. *Experimental Dermatology*, 2023.

[13] A. R. Anna Hernández Castillo. Skin lesions, what are they, types, causes, diagnosis, treatment, and more. `https://www.osmosis.org/answers/skin-lesions`. Accessed: 2nd February 2023.

[14] S. Azizi, B. Mustafa, F. Ryan, Z. Beaver, J. Freyberg, J. Deaton, A. Loh, A. Karthikesalingam, S. Kornblith, T. Chen, et al. Big self-supervised models advance medical image classification. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 3478–3488, 2021.

[15] H. Binol, A. Plotner, J. Sopkovich, B. Kaffenberger, M. K. K. Niazi, and M. N. Gurcan. Ros-net: A deep convolutional neural network for automatic identification of rosacea lesions. *Skin Research and Technology*, 26(3):413–421, 2020.

[16] M. Bonert. High magnification micrograph of extramammary paget's disease. h&e stain. `https://commons.wikimedia.org/wiki/File:Extramammary_Pagets_disease_high.jpg`. Accessed: 2nd February 2023.

[17] A. Boulemtafes, A. Derhab, and Y. Challal. A review of privacy-preserving techniques for deep learning. *Neurocomputing*, 384:21–45, 2020.

[18] A. Bożek and A. Reich. Assessment of intra-and inter-rater reliability of three methods for measuring atopic dermatitis severity: Easi, objective scorad, and iga. *Dermatology*, 233(1):16–22, 2017.

[19] T. J. Brinker, A. Hekler, A. H. Enk, J. Klode, A. Hauschild, C. Berking, B. Schilling, S. Haferkamp, D. Schadendorf, S. Fröhling, et al. A convolutional neural network trained with dermoscopic images performed on par with 145 dermatologists in a clinical melanoma image classification task. *European Journal of Cancer*, 111:148–154, 2019.

[20] T. J. Brinker, A. Hekler, A. H. Enk, J. Klode, A. Hauschild, C. Berking, B. Schilling, S. Haferkamp, D. Schadendorf, T. Holland-Letz, et al. Deep learning outperformed 136 of 157 dermatologists in a head-to-head dermoscopic melanoma image classification task. *European Journal of Cancer*, 113:47–54, 2019.

[21] T. J. Brinker, A. Hekler, C. Von Kalle, D. Schadendorf, S. Esser, C. Berking, M. T. Zacher, W. Sondermann, N. Grabe, T. Steeb, et al. Teledermatology: comparison of store-and-forward versus live interactive video conferencing. *Journal of medical Internet research*, 20(10):e11871, 2018.

[22] B. G. Buchanan and E. H. Shortliffe. Rule-based expert systems: the mycin experiments of the stanford heuristic programming project. *Computer Aided Architectural Design*, 1984.

[23] B. Burger, I. Spoerri, M. Schubert, C. Has, and P. Itin. Description of the natural course and clinical manifestations of ichthyosis with confetti caused by a novel krt10 mutation. *British journal of dermatology*, 166(2):434–439, 2012.

[24] G. Cai, Y. Zhu, Y. Wu, X. Jiang, J. Ye, and D. Yang. A multimodal transformer to fuse images and metadata for skin disease classification. *The Visual Computer*, pages 1–13, 2022.

[25] N. Cascinelli, M. Ferrario, T. Tonelli, and E. Leo. A possible new tool for clinical diagnosis of melanoma: the computer. *Journal of the American Academy of Dermatology*, 16(2):361–367, 1987.

[26] C. Cath, S. Wachter, B. Mittelstadt, M. Taddeo, and L. Floridi. Artificial intelligence and the 'good society': the us, eu, and uk approach. *Science and engineering ethics*, 24(2):505–528, 2018.

[27] K. S. Chan, Y. M. Chan, A. H. M. Tan, S. Liang, Y. T. Cho, Q. Hong, E. Yong, L. R. C. Chong, L. Zhang, G. W. L. Tan, et al. Clinical validation of an artificial intelligence-enabled wound imaging mobile application in diabetic foot ulcers. *International Wound Journal*, 19(1):114–124, 2022.

[28] C. W. Chang, M. Christian, D. H. Chang, F. Lai, T. J. Liu, Y. S. Chen, and W. J. Chen. Deep learning approach based on superpixel segmentation assisted labeling for automatic pressure ulcer diagnosis. *PloS one*, 17(2):e0264139, 2022.

[29] L. Chaves, A. Bissoto, E. Valle, and S. Avila. An evaluation of self-supervised pre-training for skin-lesion analysis. *arXiv preprint arXiv:2106.09229*, 2021.

[30] L.-C. Chen, G. Papandreou, I. Kokkinos, K. Murphy, and A. L. Yuille. Semantic image segmentation with deep convolutional nets and fully connected crfs. *arXiv preprint arXiv:1412.7062*, 2014.

[31] L.-C. Chen, G. Papandreou, I. Kokkinos, K. Murphy, and A. L. Yuille. Deeplab: Semantic image segmentation with deep convolutional nets, atrous convolution, and fully connected crfs. *IEEE transactions on pattern analysis and machine intelligence*, 40(4):834–848, 2017.

[32] L.-C. Chen, G. Papandreou, F. Schroff, and H. Adam. Rethinking atrous convolution for semantic image segmentation. *arXiv preprint arXiv:1706.05587*, 2017.

[33] L.-C. Chen, Y. Zhu, G. Papandreou, F. Schroff, and H. Adam. Encoder-decoder with atrous separable convolution for semantic image segmentation. In *Proceedings of the European conference on computer vision (ECCV)*, pages 801–818, 2018.

[34] A. Choudhury, O. Asan, et al. Role of artificial intelligence in patient safety outcomes: systematic literature review. *JMIR medical informatics*, 8(7):e18599, 2020.

[35] N. Chuchu, J. Dinnes, Y. Takwoingi, R. N. Matin, S. E. Bayliss, C. Davenport, J. F. Moreau, O. Bassett, K. Godfrey, C. O'Sullivan, et al. Teledermatology for diagnosing skin cancer in adults. *Cochrane Database of Systematic Reviews*, 2018.

[36] A. Clegg, T. Brown, D. Engels, P. Griffin, and D. Simonds. Telemedicine in a rural community hospital for remote wound care consultations. *Journal of Wound Ostomy & Continence Nursing*, 38(3):301–304, 2011.

[37] D.-A. Clevert, T. Unterthiner, and S. Hochreiter. Fast and accurate deep network learning by exponential linear units (elus). *arXiv preprint arXiv:1511.07289*, 2015.

[38] J. N. Cormier, Y. Xing, M. Ding, J. E. Lee, P. F. Mansfield, J. E. Gershenwald, M. I. Ross, and X. L. Du. Ethnic differences among patients with cutaneous melanoma. *Archives of internal medicine*, 166(17):1907–1914, 2006.

[39] F. J. Dalgard, U. Gieler, L. Tomas-Aragones, L. Lien, F. Poot, G. B. Jemec, L. Misery, C. Szabo, D. Linder, F. Sampogna, A. W. Evers, J. A. Halvorsen, F. Balieva, J. Szepietowski, D. Romanov, S. E. Marron, I. K. Altunay, A. Y. Finlay, S. S. Salek, and J. Kupfer. The psychological burden of skin diseases: A cross-sectional multicenter study among dermatological out-patients in 13 european countries. *Journal of Investigative Dermatology*, 135(4):984–991, 2015.

[40] R. Daneshjou, C. Kovarik, and J. M. Ko. Towards realization of augmented intelligence in dermatology: Advances and future directions. *arXiv preprint arXiv:2105.10477*, 2021.

[41] M. Dash, N. D. Londhe, S. Ghosh, A. Semwal, and R. S. Sonawane. Pslsnet: Automated psoriasis skin lesion segmentation using modified u-net-based fully convolutional network. *Biomedical Signal Processing and Control*, 52:226–237, 2019.

[42] T. M. de Carvalho, E. Noels, M. Wakkee, A. Udrea, and T. Nijsten. Development of smartphone apps for skin cancer risk assessment: progress and promise. *JMIR Dermatology*, 2(1):e13376, 2019.

[43] F. Decroos, S. Springenberg, T. Lang, M. Paepper, A. Zapf, D. Metze, V. Steinkraus, and A. Boeer-Auer. A deep learning approach for histopathological diagnosis of onychomycosis: not inferior to analogue diagnosis by histopathologists. *Acta Dermato-Venereologica*, 101(8):adv00532–adv00532, 2021.

[44] 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*, pages 248–255. Ieee, 2009.

[45] V. Dick, C. Sinz, M. Mittlböck, H. Kittler, and P. Tschandl. Accuracy of computer-aided diagnosis of melanoma: a meta-analysis. *JAMA dermatology*, 155(11):1291–1299, 2019.

[46] B. Dréno, D. Thiboutot, H. Gollnick, A. Y. Finlay, A. Layton, J. J. Leyden, E. Leutenegger, M. Perez, and G. A. to Improve Outcomes in Acne. Large-scale worldwide observational study of adherence with acne therapy. *International journal of dermatology*, 49(4):448–456, 2010.

[47] X. Du-Harpur, F. Watt, N. Luscombe, and M. Lynch. What is ai? applications of artificial intelligence to dermatology. *British Journal of Dermatology*, 183(3):423–430, 2020.

[48] V. Dumoulin and F. Visin. A guide to convolution arithmetic for deep learning. *arXiv preprint arXiv:1603.07285*, 2016.

[49] V. Dvorakova, R. Watson, A. Terron-Kwiatkowski, N. Andrew, and A. Irvine. Congenital reticular ichthyosiform erythroderma. *Clin Exp Dermatol*, 41:576–577, 2016.

[50] EADV 30th Congress. Burden of skin disease in europe survey, 2021.

[51] H. A. Edwards, X. Shen, and H. P. Soyer. Teledermatology adaptations in the covid-19 era. *Frontiers in medicine*, 8:674, 2021.

[52] E. J. Emanuel and R. M. Wachter. Artificial intelligence in health care: will the value match the hype? *Jama*, 321(23):2281–2282, 2019.

[53] A. Esteva, K. Chou, S. Yeung, N. Naik, A. Madani, A. Mottaghi, Y. Liu, E. Topol, J. Dean, and R. Socher. Deep learning-enabled medical computer vision. *NPJ digital medicine*, 4(1):1–9, 2021.

[54] 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*, 542(7639):115–118, 2017.

[55] F. Farnetani, M. Manfredini, J. Chester, S. Ciardo, S. Gonzalez, and G. Pellacani. Reflectance confocal microscopy in the diagnosis of pigmented macules of the face: differential diagnosis and margin definition. *Photochemical & Photobiological Sciences*, 18(5):963–969, 2019.

[56] T. Fredriksson and U. Pettersson. Severe psoriasis–oral therapy with a new retinoid. *Dermatology*, 157(4):238–244, 1978.

[57] K. Freeman, J. Dinnes, N. Chuchu, Y. Takwoingi, S. E. Bayliss, R. N. Matin, A. Jain, F. M. Walter, H. C. Williams, and J. J. Deeks. Algorithm based smartphone apps to assess risk of skin cancer in adults: systematic review of diagnostic accuracy studies. *bmj*, 368, 2020.

[58] F. Furger, L. Amruthalingam, A. Navarini, and M. Pouly. Applications of generative adversarial networks to dermatologic imaging. In *IAPR Workshop on Artificial Neural Networks in Pattern Recognition*, pages 187–199. Springer, 2020.

[59] R. S. Geiger, D. Cope, J. Ip, M. Lotosh, A. Shah, J. Weng, and R. Tang. "garbage in, garbage out" revisited: What do machine learning application papers report about human-labeled training data? *Quantitative Science Studies*, 2(3):795–827, 2021.

[60] C. Géraud and K. G. Griewank. Re: Deep learning outperformed 11 pathologists in the classification of histopathological melanoma images. *European Journal of Cancer*, 130:259–261, 2020.

[61] N. Gessert, M. Nielsen, M. Shaikh, R. Werner, and A. Schlaefer. Skin lesion classification using loss balancing and ensembles of multi-resolution efficientnets. *ISIC Challenge*, 2019.

[62] R. Girshick, J. Donahue, T. Darrell, and J. Malik. Rich feature hierarchies for accurate object detection and semantic segmentation. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 580–587, 2014.

[63] E. Goceri. Deep learning based classification of facial dermatological disorders. *Computers in Biology and Medicine*, 128:104118, 2021.

[64] T. Gong, T. Lee, C. Stephenson, V. Renduchintala, S. Padhy, A. Ndirango, G. Keskin, and O. H. Elibol. A comparison of loss weighting strategies for multi task learning in deep neural networks. *IEEE Access*, 7:141627–141632, 2019.

[65] I. Goodfellow, Y. Bengio, and A. Courville. *Deep Learning*. MIT Press, 2016.

[66] M. C. Goodier, J. G. DeKoven, J. S. Taylor, D. Sasseville, J. F. Fowler Jr, A. F. Fransway, V. A. DeLeo, J. G. Marks Jr, K. A. Zug, S. A. Hylwa, et al. Inter-rater variability in patch test readings and final interpretation using store-forward teledermatology. *Contact Dermatitis*, 85(3):274–284, 2021.

[67] M. Goyal, N. D. Reeves, A. K. Davison, S. Rajbhandari, J. Spragg, and M. H. Yap. Dfunet: Convolutional neural networks for diabetic foot ulcer classification. *IEEE Transactions on Emerging Topics in Computational Intelligence*, 4(5):728–739, 2018.

[68] J. Grob and J. Bonerandi. The 'ugly duckling'sign: identification of the common characteristics of nevi in an individual as a basis for melanoma screening. *Archives of dermatology*, 134(1):103–104, 1998.

[69] L. Guerra, A. Diociaiuti, M. El Hachem, D. Castiglia, and G. Zambruno. Ichthyosis with confetti: clinics, molecular genetics and management. *Orphanet journal of rare diseases*, 10(1):1–16, 2015.

[70] Q. Ha, B. Liu, and F. Liu. Identifying melanoma images using efficientnet ensemble: Winning solution to the siim-isic melanoma classification challenge. *arXiv preprint arXiv:2010.05351*, 2020.

[71] I. Hamzavi, H. Jain, D. McLean, J. Shapiro, H. Zeng, and H. Lui. Parametric modeling of narrowband uv-b phototherapy for vitiligo using a novel quantitative tool: the vitiligo area scoring index. *Archives of dermatology*, 140(6):677–683, 2004.

[72] S. S. Han, M. S. Kim, W. Lim, G. H. Park, I. Park, and S. E. Chang. Classification of the clinical images for benign and malignant cutaneous tumors using a deep learning algorithm. *Journal of Investigative Dermatology*, 138(7):1529–1538, 2018.

[73] S. S. Han, G. H. Park, W. Lim, M. S. Kim, J. I. Na, I. Park, and S. E. Chang. Deep neural networks show an equivalent and often superior performance to dermatologists in onychomycosis diagnosis: Automatic construction of onychomycosis datasets by region-based convolutional deep neural network. *PloS one*, 13(1):e0191493, 2018.

[74] S. N. Hart, W. Flotte, A. P. Norgan, K. K. Shah, Z. R. Buchan, T. Mounajjed, and T. J. Flotte. Classification of melanocytic lesions in selected and whole-slide images via convolutional neural networks. *Journal of Pathology Informatics*, 10, 2019.

[75] J. He, S. L. Baxter, J. Xu, J. Xu, X. Zhou, and K. Zhang. The practical implementation of artificial intelligence technologies in medicine. *Nature medicine*, 25(1):30–36, 2019.

[76] K. He, G. Gkioxari, P. Dollár, and R. Girshick. Mask r-cnn. In *Proceedings of the IEEE international conference on computer vision*, pages 2961–2969, 2017.

[77] 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*, pages 770–778, 2016.

[78] A. Hekler, J. N. Kather, E. Krieghoff-Henning, J. S. Utikal, F. Meier, F. F. Gellrich, J. Upmeier zu Belzen, L. French, J. G. Schlager, K. Ghoreschi, et al. Effects of label noise on deep learning-based skin cancer classification. *Frontiers in Medicine*, 7:177, 2020.

[79] A. Hekler, J. S. Utikal, A. H. Enk, W. Solass, M. Schmitt, J. Klode, D. Schadendorf, W. Sondermann, C. Franklin, F. Bestvater, et al. Deep learning outperformed 11 pathologists in the classification of histopathological melanoma images. *European Journal of Cancer*, 118:91–96, 2019.

[80] D. Hexsel, C. L. Hexsel, T. Dal'Forno, J. Schilling de Souza, A. F. Silva, and C. Siega. Standardized methods for photography in procedural dermatology using simple equipment. *International journal of dermatology*, 56(4):444–451, 2017.

[81] B. P. Hibler, Q. Qi, and A. M. Rossi. Current state of imaging in dermatology. *Semin Cutan Med Surg*, 35(1):2–8, 2016.

[82] K. Hornik, M. Stinchcombe, and H. White. Multilayer feedforward networks are universal approximators. *Neural networks*, 2(5):359–366, 1989.

[83] A. Hotz, V. Oji, E. Bourrat, N. Jonca, J. Mazereeuw-Hautier, R. C. Betz, U. Blume-Peytavi, K. Stieler, F. Morice-Picard, I. Schoenbuchner, et al. Expanding the clinical and genetic spectrum of krt1, krt2 and krt10 mutations in keratinopathic ichthyosis. *Acta Dermato-Venereologica*, 2016.

[84] J. Howard and S. Gugger. Fastai: a layered api for deep learning. *Information*, 11(2):108, 2020.

[85] C. Hsu, L. Amruthalingam, P. Gottfrois, and A. A. Navarini. Passion project: Pediatrics in africa — enabling wireless diagnosis for common skin diseases. `https://www.telederm.ai/`. Accessed: 2nd February 2023.

[86] J. Hu, L. Shen, and G. Sun. Squeeze-and-excitation networks. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 7132–7141, 2018.

[87] A. Jain, D. Way, V. Gupta, Y. Gao, G. de Oliveira Marinho, J. Hartford, R. Sayres, K. Kanada, C. Eng, K. Nagpal, et al. Development and assessment of an artificial intelligence–based tool for skin condition diagnosis by primary care physicians and nurse practitioners in teledermatology practices. *JAMA network open*, 4(4):e217249–e217249, 2021.

[88] S. Jiang, H. Li, and Z. Jin. A visually interpretable deep learning framework for histopathological image-based skin cancer diagnosis. *IEEE Journal of Biomedical and Health Informatics*, 25(5):1483–1494, 2021.

[89] Y. Jiang, J. Xiong, H. Li, X. Yang, W. Yu, M. Gao, X. Zhao, Y. Ma, W. Zhang, Y. Guan, et al. Recognizing basal cell carcinoma on smartphone-captured digital histopathology images with a deep neural network. *British Journal of Dermatology*, 182(3):754–762, 2020.

[90] S. John and S. Kezic. Occupational skin diseases–development and implementation of european standards on prevention of occupational skin diseases. *Journal of the European Academy of Dermatology and Venereology*, 31:3–4, 2017.

[91] M. S. Junayed, A. N. M. Sakib, N. Anjum, M. B. Islam, and A. A. Jeny. Eczemanet: A deep cnn-based eczema diseases classification. In *2020 IEEE 4th International Conference on Image Processing, Applications and Systems (IPAS)*, pages 174–179. IEEE, 2020.

[92] C. Karimkhani, R. P. Dellavalle, L. E. Coffeng, C. Flohr, R. J. Hay, S. M. Langan, E. O. Nsoesie, A. J. Ferrari, H. E. Erskine, J. I. Silverberg, et al. Global skin disease morbidity and mortality: an update from the global burden of disease study 2013. *JAMA dermatology*, 153(5):406–412, 2017.

[93] O. S. Katibi, N. C. Dlova, A. V. Chateau, and A. Mosam. The prevalence of paediatric skin conditions at a dermatology clinic in kwazulu-natal province over a 3-month period. *South African Journal of Child Health*, 10(2):121–125, 2016.

[94] A. Khan, A. Sohail, U. Zahoora, and A. S. Qureshi. A survey of the recent architectures of deep convolutional neural networks. *Artificial intelligence review*, 53(8):5455–5516, 2020.

[95] J. Kim and J. Canny. Interpretable learning for self-driving cars by visualizing causal attention. In *Proceedings of the IEEE international conference on computer vision*, pages 2942–2950, 2017.

[96] Y. J. Kim, S. S. Han, H. J. Yang, and S. E. Chang. Prospective, comparative evaluation of a deep neural network and dermoscopy in the diagnosis of onychomycosis. *PLoS One*, 15(6):e0234334, 2020.

[97] A. B. Kimball and J. S. Resneck Jr. The us dermatology workforce: a specialty remains in shortage. *Journal of the American Academy of Dermatology*, 59(5):741–745, 2008.

[98] D. P. Kingma and J. Ba. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*, 2014.

[99] S. K. Kiprono, J. W. Muchunu, and J. E. Masenga. Skin diseases in pediatric patients attending a tertiary dermatology hospital in northern tanzania: a cross-sectional study. *BMC dermatology*, 15(1):1–4, 2015.

[100] K. Leach-Kemon and A.-M. Pierre-Louis. The global burden of disease: generating evidence, guiding policy. Technical report, Institute for Health Metrics and Evaluation, University of Washington, 2013.

[101] Y. LeCun, L. Bottou, Y. Bengio, and P. Haffner. Gradient-based learning applied to document recognition. *Proceedings of the IEEE*, 86(11):2278–2324, 1998.

[102] J. J. Lee and J. C. English. Teledermatology: a review and update. *American journal of clinical dermatology*, 19(2):253–260, 2018.

[103] J. Lester, S. Taylor, and M. Chren. Under-representation of skin of colour in dermatology images: not just an educational issue. *The British Journal of Dermatology*, 180(6):1521–1522, 2019.

[104] W. Li, J. Zhuang, R. Wang, J. Zhang, and W.-S. Zheng. Fusing metadata and dermoscopy images for skin disease diagnosis. In *2020 IEEE 17th International Symposium on Biomedical Imaging (ISBI)*, pages 1996–2000. IEEE, 2020.

[105] Z. V. Lim, F. Akram, C. P. Ngo, A. A. Winarto, W. Q. Lee, K. Liang, H. H. Oon, S. T. G. Thng, and H. K. Lee. Automated grading of acne vulgaris by deep learning with convolutional neural networks. *Skin Research and Technology*, 26(2):187–192, 2020.

[106] T.-Y. Lin, P. Dollár, R. Girshick, K. He, B. Hariharan, and S. Belongie. Feature pyramid networks for object detection. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 2117–2125, 2017.

[107] T.-Y. Lin, P. Goyal, R. Girshick, K. He, and P. Dollár. Focal loss for dense object detection. In *Proceedings of the IEEE international conference on computer vision*, pages 2980–2988, 2017.

[108] Y.-L. Lin, A. Huang, C.-Y. Yang, and W.-Y. Chang. Measurement of body surface area for psoriasis using u-net models. *Computational and Mathematical Methods in Medicine*, 2022, 2022.

[109] S. Linnainmaa. *The representation of the cumulative rounding error of an algorithm as a Taylor expansion of the local rounding errors*. PhD thesis, Master's Thesis (in Finnish), Univ. Helsinki, 1970.

[110] S. Liu, Z. Chen, H. Zhou, K. He, M. Duan, Q. Zheng, P. Xiong, L. Huang, Q. Yu, G. Su, et al. Diamole: Mole detection and segmentation software for mobile phone skin images. *Journal of Healthcare Engineering*, 2021, 2021.

[111] X. Liu, S. C. Rivera, D. Moher, M. J. Calvert, and A. K. Denniston. Reporting guidelines for clinical trial reports for interventions involving artificial intelligence: the consort-ai extension. *bmj*, 370, 2020.

[112] X. Liu, L. Xie, Y. Wang, J. Zou, J. Xiong, Z. Ying, and A. V. Vasilakos. Privacy and security issues in deep learning: A survey. *IEEE Access*, 9:4566–4593, 2020.

[113] Y. Liu, A. Jain, C. Eng, D. H. Way, K. Lee, P. Bui, K. Kanada, G. de Oliveira Marinho, J. Gallegos, S. Gabriele, et al. A deep learning system for differential diagnosis of skin diseases. *Nature medicine*, 26(6):900–908, 2020.

[114] Y. Liu, Z. Ma, X. Liu, J. Liu, Z. Jiang, J. Ma, P. Yu, and K. Ren. Learn to forget: Machine unlearning via neuron masking. *arXiv preprint arXiv:2003.10933*, 2020.

[115] J. Long, E. Shelhamer, and T. Darrell. Fully convolutional networks for semantic segmentation. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 3431–3440, 2015.

[116] C. Longo, F. Farnetani, S. Ciardo, A. Cesinaro, E. Moscarella, G. Ponti, I. Zalaudek, G. Argenziano, and G. Pellacani. Is confocal microscopy a valuable tool in diagnosing nodular lesions? a study of 140 cases. *British Journal of Dermatology*, 169(1):58–67, 2013.

[117] M. Low, V. Huang, and P. Raina. Automating vitiligo skin lesion segmentation using convolutional neural networks. In *2020 IEEE 17th International Symposium on Biomedical Imaging (ISBI)*, pages 1–4. IEEE, 2020.

[118] A. N. MacLellan, E. L. Price, P. Publicover-Brouwer, K. Matheson, T. Y. Ly, S. Pasternak, N. M. Walsh, C. J. Gallant, A. Oakley, P. R. Hull, et al. The use of noninvasive imaging techniques in the diagnosis of melanoma: a prospective diagnostic accuracy study. *Journal of the American Academy of Dermatology*, 85(2):353–359, 2021.

[119] M. A. Marchetti, N. C. Codella, S. W. Dusza, D. A. Gutman, B. Helba, A. Kalloo, N. Mishra, C. Carrera, M. E. Celebi, J. L. DeFazio, et al. Results of the 2016 international skin imaging collaboration international symposium on biomedical imaging challenge: Comparison of the accuracy of computer algorithms to dermatologists for the diagnosis of melanoma from dermoscopic images. *Journal of the American Academy of Dermatology*, 78(2):270–277, 2018.

[120] M. A. Marchetti, K. Liopyris, S. W. Dusza, N. C. Codella, D. A. Gutman, B. Helba, A. Kalloo, A. C. Halpern, H. P. Soyer, C. Curiel-Lewandrowski, et al. Computer algorithms show potential for improving dermatologists' accuracy to diagnose cutaneous melanoma: Results of the international skin imaging collaboration 2017. *Journal of the American Academy of Dermatology*, 82(3):622–627, 2020.

[121] R. C. Maron, M. Weichenthal, J. S. Utikal, A. Hekler, C. Berking, A. Hauschild, A. H. Enk, S. Haferkamp, J. Klode, D. Schadendorf, et al. Systematic outperformance of 112 dermatologists in multiclass skin cancer image classification by convolutional neural networks. *European Journal of Cancer*, 119:57–65, 2019.

[122] K. Matsunaga, A. Hamada, A. Minagawa, and H. Koga. Image classification of melanoma, nevus and seborrheic keratosis by deep neural network ensemble. *arXiv preprint arXiv:1703.03108*, 2017.

[123] S. Mattessich, M. Tassavor, S. M. Swetter, and J. M. Grant-Kels. How i learned to stop worrying and love machine learning. *Clinics in dermatology*, 36(6):777–778, 2018.

[124] J. L. McAfee, A. Vij, and C. B. Warren. Store-and-forward teledermatology improves care and reduces dermatology referrals from walk-in clinics: a retrospective descriptive study. *Journal of the American Academy of Dermatology*, 82(2):499–501, 2020.

[125] A. J. McNeil, D. W. House, P. Mbala, O. Tshiani, L. Dodd, E. Cowan, Z. Chen, M. Marks, I. Saknite, E. Tkaczyk, et al. Application of u-net with inceptionv4 encoder for localizing and counting monkeypox lesions in patient photographs. Technical report, National Institute of Allergy and Infectious Diseases, 2022.

[126] A. Medela, T. Mac Carthy, S. A. A. Robles, C. M. Chiesa-Estomba, and R. Grimalt. Automatic scoring of atopic dermatitis using deep learning (ascorad): A pilot study. *JID Innovations*, page 100107, 2022.

[127] N. Meienberger, F. Anzengruber, L. Amruthalingam, R. Christen, T. Koller, J. Maul, M. Pouly, V. Djamei, and A. Navarini. Observer-independent assessment of psoriasis-affected area using machine learning. *Journal of the European Academy of Dermatology and Venereology*, 34(6):1362–1368, 2020.

[128] C. Michela and C. Matteo. Psoriasis area severity index (pasi) calculator. `http://pasi.corti.li/`. Accessed: 2nd February 2023.

[129] E. Milam and M. Leger. Use of medical photography among dermatologists: a nationwide online survey study. *Journal of the European Academy of Dermatology and Venereology*, 32(10):1804–1809, 2018.

[130] F. Milletari, N. Navab, and S.-A. Ahmadi. V-net: Fully convolutional neural networks for volumetric medical image segmentation. In *2016 fourth international conference on 3D vision (3DV)*, pages 565–571. IEEE, 2016.

[131] S. Minaee, Y. Y. Boykov, F. Porikli, A. J. Plaza, N. Kehtarnavaz, and D. Terzopoulos. Image segmentation using deep learning: A survey. *IEEE transactions on pattern analysis and machine intelligence*, 2021.

[132] M. Minsky and S. Papert. Perceptrons: An introduction to computational geometry. *Cambridge tiass., HIT*, 479:480, 1969.

[133] S. Mishra, S. Chaudhury, H. Imaizumi, and T. Yamasaki. Robustness of deep learning models in dermatological evaluation: A critical assessment. *IEICE Transactions on Information and Systems*, 104(3):419–429, 2021.

[134] M. Mohseni, J. Yap, W. Yolland, A. Koochek, and S. Atkins. Can self-training identify suspicious ugly duckling lesions? In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 1829–1836, 2021.

[135] J. A. Mojeski, M. Almashali, P. Jowdy, M. E. Fitzgerald, K. L. Brady, N. C. Zeitouni, O. R. Colegio, and G. Paragh. Ultraviolet imaging in dermatology. *Photodiagnosis and photodynamic therapy*, 30:101743, 2020.

[136] C.-I. Moon, J. Lee, H. Yoo, Y. Baek, and O. Lee. Optimization of psoriasis assessment system based on patch images. *Scientific reports*, 11(1):1–13, 2021.

[137] C. Y. Muhn, L. From, and M. Levy. Detection of artificial changes in mole size by skin self-examination. *Journal of the American Academy of Dermatology*, 42(5):754–759, 2000.

[138] C. Muñoz-López, C. Ramírez-Cornejo, M. Marchetti, S. Han, P. Del Barrio-Díaz, A. Jaque, P. Uribe, D. Majerson, M. Curi, C. Del Puerto, et al. Performance of a deep neural network in teledermatology: a single-centre prospective diagnostic study. *Journal of the European Academy of Dermatology and Venereology*, 35(2):546–553, 2021.

[139] D. H. Murphree, P. Puri, H. Shamim, S. A. Bezalel, L. A. Drage, M. Wang, M. R. Pittelkow, R. E. Carter, M. D. Davis, A. G. Bridges, et al. Deep learning for dermatologists: Part i. fundamental concepts. *Journal of the American Academy of Dermatology*, 2020.

[140] B. Mustafa, A. Loh, J. Freyberg, P. MacWilliams, M. Wilson, S. M. McKinney, M. Sieniek, J. Winkens, Y. Liu, P. Bui, et al. Supervised transfer learning at scale for medical imaging. *arXiv preprint arXiv:2101.05913*, 2021.

[141] V. Nair and G. E. Hinton. Rectified linear units improve restricted boltzmann machines. In *Icml*, 2010.

[142] C. A. Nelson, L. M. Pérez-Chada, A. Creadore, S. J. Li, K. Lo, P. Manjaly, A. B. Pournamdari, E. Tkachenko, J. S. Barbieri, J. M. Ko, et al. Patient perspectives on the use of artificial intelligence for skin cancer screening: a qualitative study. *JAMA dermatology*, 156(5):501–512, 2020.

[143] R. Niri, H. Douzi, Y. Lucas, and S. Treuillet. A superpixel-wise fully convolutional neural network approach for diabetic foot ulcer tissue classification. In *International Conference on Pattern Recognition*, pages 308–320. Springer, 2021.

[144] H. Nisar, Y. R. Tan, and Y. K. Ho. Segmentation of eczema skin lesions using u-net. In *2020 IEEE-EMBS Conference on Biomedical Engineering and Sciences (IECBES)*, pages 362–366. IEEE, 2021.

[145] A. Nozdryn-Plotnicki, J. Yap, and W. Yolland. Ensembling convolutional neural networks for skin cancer classification. *International Skin Imaging Collaboration (ISIC) Challenge on Skin Image Analysis for Melanoma Detection. MICCAI*, 2018.

[146] W. H. Organization et al. Epidemiology and management of common skin diseases in children in developing countries. Technical report, World Health Organization, 2005.

[147] V. V. Pai and R. B. Pai. Artificial intelligence in dermatology and healthcare: An overview. *Indian Journal of Dermatology, Venereology & Leprology*, 87(4), 2021.

[148] A. J. Park, J. M. Ko, and R. A. Swerlick. Crowdsourcing dermatology: Dataderm, big data analytics, and machine learning technology, 2018.

[149] P. Pasquali, S. Sonthalia, D. Moreno-Ramirez, P. Sharma, M. Agrawal, S. Gupta, D. Kumar, and D. Arora. Teledermatology and its current perspective. *Indian dermatology online journal*, 11(1):12, 2020.

[150] A. Paszke, S. Gross, F. Massa, A. Lerer, J. Bradbury, G. Chanan, T. Killeen, Z. Lin, N. Gimelshein, L. Antiga, A. Desmaison, A. Kopf, E. Yang, Z. DeVito, M. Raison, A. Tejani, S. Chilamkurthy, B. Steiner, L. Fang, J. Bai, and S. Chintala. Pytorch: An imperative style, high-performance deep learning library. In H. Wallach, H. Larochelle, A. Beygelzimer, F. d'Alché-Buc, E. Fox, and R. Garnett, editors, *Advances in Neural Information Processing Systems 32*, pages 8024–8035. Curran Associates, Inc., 2019.

[151] S. Patel, J. V. Wang, K. Motaparthi, and J. B. Lee. Artificial intelligence in dermatology for the clinician. *Clinics in dermatology*, 39(4):667–672, 2021.

[152] Y. L. Peggy Bui. Using ai to help find answers to common skin conditions. `https://blog.google/technology/health/ai-dermatology-preview-io-2021/`, 2021. Accessed: 2nd February 2023.

[153] D. A. Perednia and N. Brown. Teledermatology: one application of telemedicine. *Bulletin of the Medical Library Association*, 83(1):42, 1995.

[154] D. Prichep. Diagnostic gaps: Skin comes in many shades and so do rashes. `https://www.npr.org/sections/health-shots/2019/11/04/774910915/diagnostic-gaps-skin-comes-in-many-shades-and-so-do-rashes`, 2019. Accessed: 2nd February 2023.

[155] M. Privalov, N. Beisemann, J. E. Barbari, E. Mandelka, M. Müller, H. Syrek, P. A. Grützner, and S. Y. Vetter. Software-based method for automated segmentation and measurement of wounds on photographs using mask r-cnn: a validation study. *Journal of Digital Imaging*, 34(4):788–797, 2021.

[156] P. Puri, N. Comfere, L. A. Drage, H. Shamim, S. A. Bezalel, M. R. Pittelkow, M. D. Davis, M. Wang, A. R. Mangold, M. M. Tollefson, et al. Deep learning for dermatologists: Part ii. current applications. *Journal of the American Academy of Dermatology*, 2020.

[157] R. Raj, N. D. Londhe, and R. S. Sonawane. Deep learning based multi-segmentation for automatic estimation of psoriasis area score. In *2021 8th International Conference on Signal Processing and Integrated Networks (SPIN)*, pages 1137–1142. IEEE, 2021.

[158] P. Rajpurkar, E. Chen, O. Banerjee, and E. J. Topol. Ai in health and medicine. *Nature Medicine*, pages 1–8, 2022.

[159] P. Ramachandran, B. Zoph, and Q. V. Le. Searching for activation functions. *arXiv preprint arXiv:1710.05941*, 2017.

[160] S. Ren, K. He, R. Girshick, and J. Sun. Faster r-cnn: Towards real-time object detection with region proposal networks. *Advances in neural information processing systems*, 28, 2015.

[161] J. Resneck Jr and A. B. Kimball. The dermatology workforce shortage. *Journal of the American Academy of Dermatology*, 50(1):50–54, 2004.

[162] J. Rhodes, C. Clay, and M. Phillips. The surface area of the hand and the palm for estimating percentage of total body surface area: results of a meta-analysis. *British Journal of Dermatology*, 169(1):76–84, 2013.

[163] S. C. Rivera, X. Liu, A.-W. Chan, A. K. Denniston, and M. J. Calvert. Guidelines for clinical trial protocols for interventions involving artificial intelligence: the spirit-ai extension. *Bmj*, 370, 2020.

[164] O. Ronneberger, P. Fischer, and T. Brox. U-net: Convolutional networks for biomedical image segmentation. In *International Conference on Medical image computing and computer-assisted intervention*, pages 234–241. Springer, 2015.

[165] B. Rostami, D. Anisuzzaman, C. Wang, S. Gopalakrishnan, J. Niezgoda, and Z. Yu. Multiclass wound image classification using an ensemble deep cnn-based classifier. *Computers in Biology and Medicine*, 134:104536, 2021.

[166] V. Rotemberg, N. Kurtansky, B. Betz-Stablein, L. Caffery, E. Chousakos, N. Codella, M. Combalia, S. Dusza, P. Guitera, D. Gutman, et al. A patient-centric dataset of images and metadata for identifying melanomas using clinical context. *Scientific data*, 8(1):1–8, 2021.

[167] A. G. Roy, J. Ren, S. Azizi, A. Loh, V. Natarajan, B. Mustafa, N. Pawlowski, J. Freyberg, Y. Liu, Z. Beaver, et al. Does your dermatology classifier know what it doesn't know? detecting the long-tail of unseen conditions. *Medical Image Analysis*, 75:102274, 2022.

[168] D. E. Rumelhart, G. E. Hinton, and R. J. Williams. Learning internal representations by error propagation. Technical report, California Univ San Diego La Jolla Inst for Cognitive Science, 1985.

[169] S. J. Russell and P. Norvig. *Artificial Intelligence: a modern approach*. Pearson, 4 edition, 2020.

[170] S. Sabour. Comment on "the use of noninvasive imaging techniques in the diagnosis of melanoma: A prospective diagnostic accuracy study". *Journal of the American Academy of Dermatology*, 85(2):e85, 2021.

[171] B. Saka, E. Lauressergues, G. Mahamadou, L. Matel, C. Abilogoun, H. Adégbidi, K. Ahogo, A. Akakpo, C. Akakpo, E. Akata, et al. Deuxièmes assises de télédermatologie africaines—lomé (togo). *La Presse Médicale Formation*, 1(2):198–202, 2020.

[172] T. Sangers, T. Nijsten, and M. Wakkee. Mobile health skin cancer risk assessment campaign using artificial intelligence on a population-wide scale: A retrospective cohort analysis. *Journal of the European Academy of Dermatology and Venereology: JEADV*, 2021.

[173] K. V. Sarma, S. Harmon, T. Sanford, H. R. Roth, Z. Xu, J. Tetreault, D. Xu, M. G. Flores, A. G. Raman, R. Kulkarni, et al. Federated learning improves site performance in multicenter deep learning without data sharing. *Journal of the American Medical Informatics Association*, 28(6):1259–1264, 2021.

[174] M. J. Schaap, N. J. Cardozo, A. Patel, E. M. de Jong, B. van Ginneken, and M. M. Seyger. Image-based automated psoriasis area severity index scoring by convolutional neural networks. *Journal of the European Academy of Dermatology and Venereology*, 36(1):68–75, 2022.

[175] D. Schiff and J. Borenstein. How should clinicians communicate with patients about the roles of artificially intelligent team members? *AMA Journal of Ethics*, 21(2):138–145, 2019.

[176] J. Schroeter, C. Myers-Colet, D. Arnold, and T. Arbel. Segmentation-consistent probabilistic lesion counting. In *Medical Imaging with Deep Learning*, 2022.

[177] L. Schuchter, D. J. Schultz, M. Synnestvedt, B. J. Trock, D. Guerry, D. E. Elder, R. Elenitsas, W. H. Clark, and A. Halpern. A prognostic model for predicting 10-year survival in patients with primary melanoma. *Annals of internal medicine*, 125(5):369–375, 1996.

[178] S. Seité, A. Khammari, M. Benzaquen, D. Moyal, and B. Dréno. Development and accuracy of an artificial intelligence algorithm for acne grading from smartphone photographs. *Experimental dermatology*, 28(11):1252–1257, 2019.

[179] R. R. Selvaraju, M. Cogswell, A. Das, R. Vedantam, D. Parikh, and D. Batra. Grad-cam: Visual explanations from deep networks via gradient-based localization. In *Proceedings of the IEEE international conference on computer vision*, pages 618–626, 2017.

[180] C. Shorten and T. M. Khoshgoftaar. A survey on image data augmentation for deep learning. *Journal of big data*, 6(1):1–48, 2019.

[181] K. Simonyan, A. Vedaldi, and A. Zisserman. Deep inside convolutional networks: Visualising image classification models and saliency maps. In *In Workshop at International Conference on Learning Representations*. Citeseer, 2014.

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

[183] A. Singh, S. Sengupta, and V. Lakshminarayanan. Explainable deep learning models in medical image analysis. *Journal of Imaging*, 6(6):52, 2020.

[184] L. N. Smith. A disciplined approach to neural network hyper-parameters: Part 1–learning rate, batch size, momentum, and weight decay. *arXiv preprint arXiv:1803.09820*, 2018.

[185] A. Stratigos, A. Forsea, R. Van Der Leest, E. De Vries, E. Nagore, J.-L. Bulliard, M. Trakatelli, J. Paoli, K. Peris, J. Hercogova, et al. Euromelanoma: a dermatology-led european campaign against nonmelanoma skin cancer and cutaneous melanoma. past, present and future. *British Journal of Dermatology*, 167:99–104, 2012.

[186] I. Sutskever, J. Martens, G. Dahl, and G. Hinton. On the importance of initialization and momentum in deep learning. In *International conference on machine learning*, pages 1139–1147. PMLR, 2013.

[187] C. Szegedy, S. Ioffe, V. Vanhoucke, and A. A. Alemi. Inception-v4, inception-resnet and the impact of residual connections on learning. In *Thirty-first AAAI conference on artificial intelligence*, 2017.

[188] C. Szegedy, W. Liu, Y. Jia, P. Sermanet, S. Reed, D. Anguelov, D. Erhan, V. Vanhoucke, and A. Rabinovich. Going deeper with convolutions. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 1–9, 2015.

[189] 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*, pages 2818–2826, 2016.

[190] M. Tan and Q. Le. Efficientnet: Rethinking model scaling for convolutional neural networks. In *International conference on machine learning*, pages 6105–6114. PMLR, 2019.

[191] E. Tensen, J. Van Der Heijden, M. Jaspers, and L. Witkamp. Two decades of teledermatology: current status and integration in national healthcare systems. *Current dermatology reports*, 5(2):96–104, 2016.

[192] The American Medical Association. Augmented intelligence in health care. `https://www.ama-assn.org/system/files/2019-01/augmented-intelligence-policy-report.pdf`, 2019. Accessed: 2nd February 2023.

[193] B. X. Tran, G. T. Vu, G. H. Ha, Q.-H. Vuong, M.-T. Ho, T.-T. Vuong, V.-P. La, M.-T. Ho, K.-C. P. Nghiem, H. L. T. Nguyen, et al. Global evolution of research in artificial intelligence in health and medicine: a bibliometric study. *Journal of clinical medicine*, 8(3):360, 2019.

[194] P. Tschandl, N. Codella, B. N. Akay, G. Argenziano, R. P. Braun, H. Cabo, D. Gutman, A. Halpern, B. Helba, R. Hofmann-Wellenhof, et al. Comparison of the accuracy of human readers versus machine-learning algorithms for pigmented skin lesion classification: an open, web-based, international, diagnostic study. *The lancet oncology*, 20(7):938–947, 2019.

[195] P. Tschandl, C. Rinner, Z. Apalla, G. Argenziano, N. Codella, A. Halpern, M. Janda, A. Lallas, C. Longo, J. Malvehy, et al. Human–computer collaboration for skin cancer recognition. *Nature Medicine*, 26(8):1229–1234, 2020.

[196] P. Tschandl, C. Rosendahl, B. N. Akay, G. Argenziano, A. Blum, R. P. Braun, H. Cabo, J.-Y. Gourhant, J. Kreusch, A. Lallas, et al. Expert-level diagnosis of nonpigmented skin cancer by combined convolutional neural networks. *JAMA dermatology*, 155(1):58–65, 2019.

[197] Ç. Turan, N. Metin, Z. Utlu, Ü. Öner, and Ö. S. Kotan. Change of the diagnostic distribution in applicants to dermatology after covid-19 pandemic: What it whispers to us? *Dermatologic Therapy*, 33(4):e13804, 2020.

[198] A. Udrea, G. Mitra, D. Costea, E. Noels, M. Wakkee, D. Siegel, T. de Carvalho, and T. Nijsten. Accuracy of a smartphone application for triage of skin lesions based on machine learning algorithms. *Journal of the European Academy of Dermatology and Venereology*, 34(3):648–655, 2020.

[199] US Food & Drug Administration. Artificial intelligence and machine learning in software as a medical device. `https://www.fda.gov/medical-devices/software-medical-device-samd/artificial-intelligence-and-machine-learning-software-medical-device`. Accessed: 2nd February 2023.

[200] J. Van der Heijden, N. De Keizer, J. Bos, P. Spuls, and L. Witkamp. Teledermatology applied following patient selection by general practitioners in daily practice improves efficiency and quality of care at lower cost. *British Journal of Dermatology*, 165(5):1058–1065, 2011.

[201] B. H. van der Velden, H. J. Kuijf, K. G. Gilhuijs, and M. A. Viergever. Explainable artificial intelligence (xai) in deep learning-based medical image analysis. *arXiv preprint arXiv:2107.10912*, 2021.

[202] M. C. van Zon, J. D. van der Waa, M. Veta, and G. A. Krekels. Whole-slide margin control through deep learning in mohs micrographic surgery for basal cell carcinoma. *Experimental Dermatology*, 30(5):733–738, 2021.

[203] F. Veronese, F. Branciforti, E. Zavattaro, V. Tarantino, V. Romano, K. M. Meiburger, M. Salvi, S. Seoni, and P. Savoia. The role in teledermoscopy of an inexpensive and easy-to-use smartphone device for the classification of three types of skin lesions using convolutional neural networks. *Diagnostics*, 11(3):451, 2021.

[204] M. Vestergaard, P. Macaskill, P. Holt, and S. Menzies. Dermoscopy compared with naked eye examination for the diagnosis of primary melanoma: a meta-analysis of studies performed in a clinical setting. *British Journal of Dermatology*, 159(3):669–676, 2008.

[205] S. Wachter, B. Mittelstadt, and L. Floridi. Why a right to explanation of automated decision-making does not exist in the general data protection regulation. *International Data Privacy Law*, 7(2):76–99, 2017.

[206] F. Wang, R. Kaushal, and D. Khullar. Should health care demand interpretable artificial intelligence or accept "black box" medicine?, 2020.

[207] M. Z. Wang, N. I. Comfere, and D. H. Murphree. Deep learning for automating the organization of institutional dermatology image stores. In *2019 41st Annual International Conference of the IEEE Engineering in Medicine and Biology Society (EMBC)*, pages 4479–4482. IEEE, 2019.

[208] E. M. Warshaw, Y. J. Hillman, N. L. Greer, E. M. Hagel, R. MacDonald, I. R. Rutks, and T. J. Wilt. Teledermatology for diagnosis and management of skin conditions: a systematic review. *Journal of the American Academy of Dermatology*, 64(4):759–772, 2011.

[209] J. K. Winkler, C. Fink, F. Toberer, A. Enk, T. Deinlein, R. Hofmann-Wellenhof, L. Thomas, A. Lallas, A. Blum, W. Stolz, et al. Association between surgical skin markings in dermoscopic images and diagnostic performance of a deep learning convolutional neural network for melanoma recognition. *JAMA dermatology*, 155(10):1135–1141, 2019.

[210] K. Włodarek, A. Stefaniak, Ł. Matusiak, and J. C. Szepietowski. Could residents adequately assess the severity of hidradenitis suppurativa? interrater and intrarater reliability assessment of major scoring systems. *Dermatology*, 236(1):8–14, 2020.

[211] A. Woodley. Can teledermatology meet the needs of the remote and rural population? *British Journal of Nursing*, 30(10):574–579, 2021.

[212] H. Wu, H. Yin, H. Chen, M. Sun, X. Liu, Y. Yu, Y. Tang, H. Long, B. Zhang, J. Zhang, et al. A deep learning, image based approach for automated diagnosis for inflammatory skin diseases. *Annals of translational medicine*, 8(9), 2020.

[213] P. Xie, T. Li, J. Liu, F. Li, J. Zhou, and K. Zuo. Analyze skin histopathology images using multiple deep learning methods. In *2021 IEEE 3rd International Conference on Frontiers Technology of Information and Computer (ICFTIC)*, pages 374–377. IEEE, 2021.

[214] S. Xie, R. Girshick, P. Dollár, Z. Tu, and K. He. Aggregated residual transformations for deep neural networks. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 1492–1500, 2017.

[215] R. Yamashita, M. Nishio, R. K. G. Do, and K. Togashi. Convolutional neural networks: an overview and application in radiology. *Insights into imaging*, 9(4):611–629, 2018.

[216] Y. Yang, L. Guo, Q. Wu, M. Zhang, R. Zeng, H. Ding, H. Zheng, J. Xie, Y. Li, Y. Ge, et al. Construction and evaluation of a deep learning model for assessing acne vulgaris using clinical images. *Dermatology and therapy*, 11(4):1239–1248, 2021.

[217] Y. Yang, J. Wang, F. Xie, J. Liu, C. Shu, Y. Wang, Y. Zheng, and H. Zhang. A convolutional neural network trained with dermoscopic images of psoriasis performed on par with 230 dermatologists. *Computers in Biology and Medicine*, 139:104924, 2021.

[218] J. Yap, W. Yolland, and P. Tschandl. Multimodal skin lesion classification using deep learning. *Experimental dermatology*, 27(11):1261–1267, 2018.

[219] S. W. Youn, C. W. Choi, B. R. Kim, and J. B. Chae. Reduction of inter-rater and intra-rater variability in psoriasis area and severity index assessment by photographic training. *Annals of Dermatology*, 27(5):557–562, 2015.

[220] 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*, 36(4):994–1004, 2016.

[221] G. A. Zakhem, J. W. Fakhoury, C. C. Motosko, and R. S. Ho. Characterizing the role of dermatologists in developing artificial intelligence for assessment of skin cancer. *Journal of the American Academy of Dermatology*, 85(6):1544–1556, 2021.

[222] A. Zhang, Z. C. Lipton, M. Li, and A. J. Smola. Dive into deep learning. *arXiv preprint arXiv:2106.11342*, 2021.

[223] B. Zhang and A. Dafoe. Artificial intelligence: American attitudes and trends. *Available at SSRN 3312874*, 2019.

[224] C.-Y. Zhu, Y.-K. Wang, H.-P. Chen, K.-L. Gao, C. Shu, J.-C. Wang, L.-F. Yan, Y.-G. Yang, F.-Y. Xie, and J. Liu. A deep learning based framework for diagnosing multiple skin diseases in a clinical environment. *Frontiers in medicine*, 8, 2021.

[225] X. Zhu, B. Zheng, W. Cai, J. Zhang, S. Lu, X. Li, L. Xi, and Y. Kong. Deep learning-based diagnosis models for onychomycosis in dermoscopy. *Mycoses*, 2022.

Publications

Peer-reviewed Publications

- (research article, accepted) L. Amruthalingam, N. Mang, P. Gottfrois, A. Gonzalez Jimenez, M. Kunz, T. Koller, M. Pouly and A. A. Navarini, "Objective Hand Eczema Severity Assessment with Automated Lesion Anatomical Stratification", Journal of Experimental Dermatology, 2023
- (research article, accepted) L. Amruthalingam, O. Buerzle, P. Gottfrois, A. Gonzalez Jimenez, A. Roth, T. Koller, M. Pouly and A. A. Navarini, "Quantification of Efflorescences in Pustular Psoriasis using Deep Learning", Journal of Healthcare Informatics Research, 2022
- (research article, accepted) L. Amruthalingam, P. Gottfrois, A. Gonzalez Jimenez, B. Gökduman, M. Kunz, T. Koller, DERMANATOMY Consortium, M. Pouly and A. A. Navarini, "Improved Diagnosis by Automated Macro- and Micro-anatomical Region Mapping of Skin Photographs", Journal of the European Academy of Dermatology and Venereology, 2022
- (conference paper, accepted) F. Groeger, P. Gottfrois, L. Amruthalingam, A. Gonzalez Jimenez, S. Lionetti, A. A. Navarini and M. Pouly, "Towards Reducing The Need For Annotations In Digital Dermatology With Self-Supervised Learning ", ECAI Workshop on Scarce Data in Artificial Intelligence for Healthcare, 2022
- (conference paper, accepted) A. Gonzalez Jimenez, S. Lionetti, L. Amruthalingam, P. Gottfrois, M. Pouly and A. A. Navarini, "SANO: Score-based Anomaly Localization for Dermatology", ECAI Workshop on Scarce Data in Artificial Intelligence for Healthcare, 2022
- (short conference paper, to be resubmitted) L. Amruthalingam, P. Gottfrois, A. Gonzalez Jimenez, F. Rapelanoro Rabenja, I. Traoré, M. Pouly and A. A. Navarini, "Medical Image Collection in Sub-Saharan Africa", conference to be defined, 2022
- (research article, accepted) N. Meienberger, F. Anzengruber, L. Amruthalingam, R. Christen, T. Koller, J. T. Maul, M. Pouly, V. Djamei, A. A. Navarini, "Observer-independent assessment of psoriasis-affected area using machine learning", Journal of the European Academy of Dermatology and Venereology, 2020

- (conference paper, accepted) F. Furger, L. Amruthalingam, A. A. Navarini and M. Pouly, “Applications of Generative Adversarial Networks to Dermatologic Imaging”, IAPR Workshop on Artificial Neural Networks in Pattern Recognition, 2020

Posters

- L. Amruthalingam, O. Buerzle, P. Gottfrois, A. Gonzalez Jimenez, A. Roth, T. Koller, M. Pouly and A. A. Navarini, “Automated Severity Grading of Palmoplantar Pustular Psoriasis”, European Academy of Dermatology and Venereology, DBE Research Day, 2021
- L. Amruthalingam, P. Gottfrois, H. Andriambololoniaina, S. Zhao, R. Philemon, V. Amann, Y. Xiao, H. P. Soyer, P. Pasquali, L. Caffery, H. Hassan, N. Doss, F. Dassoni, M. A. Rodrigues, F. Rapelanoro Rabenja, P. Schmid-Grendelmeier, W. Huang, D. Mavura, X. Chen, A. A. Navarini, C. Hsu, “Artificial Intelligence in Dermatology: an Application in Sub-Saharan Africa, American Academy of Dermatology Annual Meeting, 2021
- P. Gottfrois, L. Amruthalingam, T. Koller, M. Pouly and A. A. Navarini, “Skin Lesion Classification”, DBE Research Day, 2019
- L. Amruthalingam, T. Koller, M. Pouly and A. A. Navarini, “Deep Learning in Clinical Dermatology”, DBE Summer School, 2019

Talks

- L. Amruthalingam, P. Gottfrois, A. Gonzalez Jimenez, T. Koller, M. Pouly and A. A. Navarini, “Automated Anatomical Mapping of Hand Eczema”, Universitätsspital Basel, 2021
- L. Amruthalingam, P. Gottfrois, T. Koller, M. Pouly and A. A. Navarini, “Automated Severity Grading of Palmoplantar Pustular Psoriasis”, Swiss Society for Dermatology and Venereology, 2021
- L. Amruthalingam, P. Gottfrois, T. Koller, M. Pouly and A. A. Navarini, “Localization of Skin Dermatology Pictures on the Body”, Swiss Society for Dermatology and Venereology, 2020
- F. Furger, L. Amruthalingam, A. A. Navarini and M. Pouly, “Applications of Generative Adversarial Networks to Dermatologic Imaging”, IAPR Workshop on Artificial Neural Networks in Pattern Recognition, 2020
- L. Amruthalingam, T. Koller, M. Pouly and A. A. Navarini, “Skin Image Localization on the Human Body”, Roche Digital Day, 2020

- L. Amruthalingam, T. Koller, M. Pouly and A. A. Navarini, “Applications of Artificial Intelligence in Dermatology”, Middle East International Dermatology and Aesthetic Medicine, 2020
- L. Amruthalingam, T. Koller, M. Pouly and A. A. Navarini, “Artificial Intelligence for Clinical Dermatology”, European Academy of Dermatology and Venereology, 2019