Explainable Neural Networks for Skin Lesion Diagnosis

Jalaboi, Raluca Alexandra

Publication date:
2023

Document Version
Publisher's PDF, also known as Version of record

Citation (APA):
Jalaboi, R. A. (2023). *Explainable Neural Networks for Skin Lesion Diagnosis*. Technical University of Denmark.

Doctor of Philosophy
Doctoral thesis in Computer Science

Explainable Neural Networks for Skin Lesion Diagnosis

Author: Raluca Jalaboi

Supervisor: Prof. Ole Winther

Co-supervisors: Dr. Alfiia Galimzianova

DTU Compute
Department of Applied Mathematics and Computer Science
Technical University of Denmark
Richard Petersens Plads
Building 324
2800 Kongens Lyngby, Denmark

Summary (English)

There are more than 3,000 skin diseases, and more than a third of the world's population will suffer from at least one skin disease throughout their lives. With general practitioners obtaining only a 61% dermatological diagnosis accuracy, automation is an intuitive next step to lower the burden on medical professionals and decrease the time to diagnosis and treatment for patients. Despite automated diagnosis methods achieving expert-level performance, adoption is hindered by their limited ability to explain their reasoning. Additionally, people have been shown to prefer clear, concise explanations tailored to their understanding of a field.

Multiple explainability methods have been proposed, but their results were primarily tested qualitatively by machine learning researchers. Quantitative analyses are rarely performed due to their reliance on domain experts, which have low availability and are expensive to hire as consultants.

In this thesis, we propose a novel method for intrinsically explainable convolutional neural networks (ConvNets) that achieves expert-level explainability at no cost to the classification performance.

First, we investigate what explainability means to experts through an explainability dataset for diagnosing six skin diseases, consisting of diagnoses and explanations from eight board-certified dermatologists. The dataset shows dermatologists may explain their decision differently even when agreeing on the diagnosis.

Using this data, we benchmark a representative set of state-of-the-art ConvNet architectures used for skin lesion diagnosis. We find that despite the ability of current explainability methods to produce explanations, more research is required to achieve specialist-level performance.

To move closer to this goal, we introduce two intrinsically explainable ConvNet architectures trained to emulate a dermatologist's decision process. The two ConvNets achieve almost expert-level explainability at no cost to the diagnosis performance.

Finally, we prove that our method can be applied in other domains by developing an image quality assessment network within a teledermatological context. We obtained similar results to the skin disease diagnosis networks and reduced the number of low-quality images sent to the dermatologists by 70%.

Resume (Dansk)

Der findes mere end 3000 hudsygdomme, og over en tredjedel af verdens befolkning vil lide af mindst én hudsygdom i løbet af deres liv. Eftersom alment praktiserende læger kun har en diagnostisk nøjagtighed på 61% inden for dermatologi, er automatisering ved hjælp af neurale netværk det næste nærliggende skridt for at understøtte læger og andet sundhedspersonale i deres arbejde og hjælpe patienter til hurtigere at opnå den korrekte diagnose og behandling. På trods af at automatiserede diagnosemetoder kan præstere en nøjagtighed på samme niveau som specialister, er udbredelse og anvendelse af disse værktøjer begrænset af deres manglende evne til at forklare de processer, der fører til en bestemt klassificering - såsom hvorfor en diagnose er valgt frem for en anden. Derudover fremgår det af studier, at mennesker generelt foretrækker klare, kortfattede forklaringer tilpasset deres forståelsesramme inden for et givent felt.

En række forklaringsmodeller er blevet undersøgt for at adressere denne udfordring, men de er primært blevet testet kvalitativt af forskere inden for maskinlæring. Kvantitative analyser er kun sjældent blevet udført, da de er afhængige af eksperter inden for det givne domæne, og disse er ofte svært tilgængelige og dyre at hyre som konsulenter.

I denne specialeafhandling fremlægger vi en ny metode til at opnå konvolutionelle neurale netværk (ConvNets), som kan forklare deres beslutningsprocesser på samme niveau som specialister, uden at der sker på bekostning af klassificeringernes nøjagtighed.

Først undersøgte vi, hvordan specialister begrunder deres beslutninger ved hjælp af forklarlige dataset bestående af diagnoser af seks hudsygdomme samt begrundelser for disse diagnoser, som er foretaget af otte autoriserede dermatologer. Af disse data fremgår det, at dermatologernes begrundelse for en diagnose kan være forskellig - også i tilfælde hvor de er enige om diagnosen.

På baggrund af disse data kunne vi benchmarke et repræsentativt udsnit af andre ConvNet-arkitekturer for diagnose af hudlæsioner. På trods af at disse netværk også kunne forklare den valgte diagnose, kunne de ikke gøre det på samme niveau som den menneskelige pendant. Derfor introducerer vi to forklarlige ConvNet-arkitekturer, som er trænet til at efterligne dermatologers beslutningsprocesser. De to ConvNets kunne forklare deres beslutningsproces på næsten samme niveau som specialister, og uden at det skete på bekostning af den diagnostiske nøjagtighed.

Slutteligt kan vi påvise, at vores metode kan benyttes inden for andre områder inden for teledermatologi ved at skabe et netværk for evaluering af billedkvalitet. Her opnåede vi lignende resultater som med netværket for huddiagnose, hvor vi reducerede antallet af billeder i dårlig kvalitet, som blev sendt til dermatologerne, med 70%.

Preface

This thesis was prepared at the Cognitive Systems section of DTU Compute, Department of Applied Mathematics and Computer Science at the Technical University of Denmark. It constitutes a partial fulfillment of the requirements for acquiring a Ph.D. degree at the Technical University of Denmark.

This thesis was supervised by Ole Winther at DTU and Alfiia Galimzianova at Medable. The Ph.D. project was sponsored by Omhu/Medable and the Innovation Fund Denmark, grant number 0153-00154A. The work was carried out at DTU, Omhu/Medable from October 2020 to July 2023, with the exception of an external research visit at the Imperial College London in April 2023.

The thesis consists of three research papers (two peer-reviewed and one pre-print) on the topic of explainability for automated skin lesion diagnosis.

Kongens Lyngby, 22$^{\text{nd}}$ August 2023

Raluca Jalaboi

Acknowledgements

First and foremost, I would like to thank my supervisors, Alfiia Galimzianova and Ole Winther, for their guidance, support, and encouragement over these three years. I would also like to thank my external research stay supervisor Ben Glocker for the inspiring collaboration and exciting discussions.

Additionally, I would like to extend my gratitude towards my collaborators Frederik Faye, Dan Jørgensen, Mauricio Orbes, and Kenneth Thomsen for their fantastic work and great ideas. I would also like to thank all my coworkers from Omhu that supported me during this time, especially the fantastic team of dermatologists that helped build the dataset. A big thanks to Martin Holm Jensen and Christian Sejersen for offering me the opportunity to embark on this journey and to the Innovation Fund Denmark for sponsoring this project.

Finally, a huge and heartfelt thank you to my boyfriend, Bogdan Spiridon, for his support, patience, and trust in me.

List of Publications

Contributions included in this thesis:

[Jalaboi et al., 2023a]. Raluca Jalaboi, Frederik Faye, Mauricio Orbes-Arteaga, Dan Jørgensen, Ole Winther, and Alfiia Galimzianova. ”DermX: An end-to-end framework for explainable automated dermatological diagnosis”. In *Medical Image Analysis* 83 (2023): 102647.

[Jalaboi et al., 2023b]. Raluca Jalaboi, Ole Winther, and Alfiia Galimzianova. ”Dermatological Diagnosis Explainability Benchmark for Convolutional Neural Networks”. *arXiv preprint* arXiv:2302.12084 (2023).

[Jalaboi et al., 2023c]. Raluca Jalaboi, Ole Winther, and Alfiia Galimzianova. ”Explainable image quality assessments in teledermatological photography.” *Telemedicine and e-Health (2023).*

Other contributions:

[Jalaboi et al., 2021] Raluca Jalaboi, Mauricio Orbes Arteaga, Dan Richter Jørgensen, Ionela Manole, Oana Ionescu Bozdog, Andrei Chiriac, Ole Winther, and Alfiia Galimzianova. ”Explainability of Convolutional Neural Networks for Dermatological Diagnosis.” In *Proceedings of 9th World Congress of Teledermatology, Imaging and AI for Skin Diseases 2021* 7, no. 1 (2021): e35437.

[Thomsen et al., 2023] Kenneth Thomsen, Raluca Jalaboi, Ole Winther, Hans Bredsted Lomholt, Henrik F. Lorentzen, Trine Høgsberg, Henrik Egekvist, Lene Hedelund, Sofie Jørgensen, Sanne Frost, and Lars Iversen. ”Physician level assessment of hirsute women and eligibility for laser treatment with Deep Learning.” *Unpublished.*

Contents

CHAPTER 1

Introduction

More than 30% of the world's population will at some point in their lives experience at least one skin disease [Hay et al., 2014] out of the approximately 3,000 skin diseases that have been identified so far [Lewin Group, 2005]. Yet, the average general practitioner diagnoses a skin condition with an accuracy of around 60% [Federman et al., 1999], leading to inadequate triaging, treatment plans, and patient outcomes. Thus, there is a pressing need for reliable automated tools to improve the disease management process for patients suffering from dermatological conditions.

Deep learning has been successfully applied in a wide range of medical applications for different modalities and tasks [Shehab et al., 2022], with convolutional neural networks (ConvNets) [LeCun et al., 1998] currently being the preferred approach in medical imaging [Castiglioni et al., 2021]. Within dermatology, Esteva et al. [2017] ignited the exploration of deep learning methods with their seminal work on skin cancer diagnosis which achieved expert-level performance. Since then, skin lesion diagnosis has become one of the primary research areas in medical image analysis [Thomsen et al., 2020]. However, despite the consistent advances and proof of expert-level performance [Tschandl et al., 2018, 2020, Roy et al., 2022], such automated methods have not yet been widely adopted by healthcare systems. One of the main hurdles detracting from adoption is the limited explainability of a ConvNet's decision process [Kelly et al., 2019].

Multiple explainability approaches that can offer insights into the ConvNet decision mechanisms have been investigated [Molnar, 2020, Jin et al., 2022]. In particular, backpropagation-based methods stand out as a staple of medical imaging due to their ease of implementation and low computational requirements [Van der Velden et al., 2022]. One downside of these explanations is the researchers' limited capability for performing quantitative evaluations.

Thorough, quantitative evaluations require domain expert involvement, which is an expensive and time-consuming endeavor. Because of the high costs, limited effort has gone into quantitatively evaluating explanations despite the possibility of generating novel insights into ConvNet decision mechanisms [Tschandl et al., 2020]. Rajpurkar et al. [2018] show that explanations can improve the clinicians' performance when using an automated method as a second opinion, while González-Gonzalo et al. [2022] underline the importance of explanations in creating trustworthy automated methods. Explainability is thus a core component to ensuring that automation will be incorporated into the healthcare system.

To enable the healthcare system's much-needed adoption of automated methods, we need to create explainable diagnosis ConvNets whose explainability can be objectively compared to that of dermatologists.

1.1 Thesis structure

In this thesis, we explore skin disease diagnosis explainability from a dermatologist's perspective and propose two intrinsically explainable architectures for dermatological diagnosis. Then, we objectively benchmark the explainability of ConvNet architectures commonly used for this task. Finally, we prove that this methodology can be applied with similar results in image quality assessment for teledermatology.

Chapter 2 introduces the key concepts used in the rest of the thesis. First, we present dermatological concepts fundamental to our research, and afterward we explore explanations from a social sciences perspective. Second, we describe state-of-the-art explainability methods for ConvNets and compare their approaches to explanations. Finally, we summarize the main trends in ConvNets for dermatological applications.

Chapter 3 summarizes the end-to-end methodology for creating explainable dermatological diagnosis ConvNets described in Jalaboi et al. [2023a]. We first introduce DermXDB, a skin disease explainability dataset which serves as a reference standard for further research into explainability. Then, we introduce DermX and DermX+, two intrinsically explainable ConvNet architectures for skin lesion diagnosis.

Chapter 4 presents a ConvNet architecture explainability benchmark that compares ConvNet explainability maps created using gradient class activation maps with explanations maps derived from the dataset introduced in the previous chapter. This chapter is based on Jalaboi et al. [2023b].

Chapter 5 illustrates the versatility of the DermX methodology by applying it to image quality assessments within a teledermatological environment. This work was presented in Jalaboi et al. [2023c].

Finally, Chapter 6 summarizes the contributions of this thesis, highlights their impact, and suggests some directions for future work.

CHAPTER 2

Background

This chapter introduces the key concepts used in this thesis. First, we introduce fundamental dermatological concepts for skin lesion diagnosis in Section 2.1. Section 2.2 discusses explainability from a social sciences perspective, focusing on how people expect and interpret explanations. Afterward, we discuss the main explanation methods proposed for ConvNets in Section 2.3 and compare them in Section 2.4. Finally, Section 2.5 creates an overview of ConvNets in dermatological applications.

2.1 A short introduction to dermatological diagnosis

Dermatological diagnosis is a difficult field, with more than 3,000 different diseases having been identified [Lewin Group, 2005], and up to a third of the world's population reporting at least one skin condition during their lives [Hay et al., 2014]. The large number of potential patients, the limited number of dermatologists, the ease with which symptoms can be observed, and the ubiquitousness of smartphone cameras promote dermatological diagnosis as a prime candidate for automation.

Diagnosing skin conditions is a complex task that generally requires both a thorough skin examination, as well as knowledge about the patient's history (referred to as anamnesis), the lesion's history, and the patient's non-visual symptoms. The anamnesis focuses on the patient's history of disease and recent events or experiences that may have triggered a skin reaction, e.g. travel to tropical countries or exposure to poisonous plants. The lesion history provides information on the disease progression or malignancy. The skin examination focuses on the lesion morphology, palpation and texture, color, and configuration and distribution [Fitzpatrick and High, 2017]. Although all four components are necessary in many cases to arrive at a diagnosis, some diseases have a distinctive appearance that enables healthcare professionals to diagnose them from morphological traits alone [Oakley, 2017].

Morphological attributes are often grouped into basic and additional terms [Nast et al., 2016], as illustrated in Figure 2.1 and Figure 2.2, respectively. Additional descriptive terms are often used alongside basic terms to describe the lesion more comprehensively, which may help differentiate between diseases. For example, the color of the scales present in lesions can help with differentiating between seborrheic dermatitis and psoriasis: greasy, yellow scales are characteristic of seborrheic dermatitis, while silvery-white scales are a hallmark of psoriasis [Oakley, 2017]. Other additional descriptive terms include texture, shape, or how well-defined the edges of a lesion are.

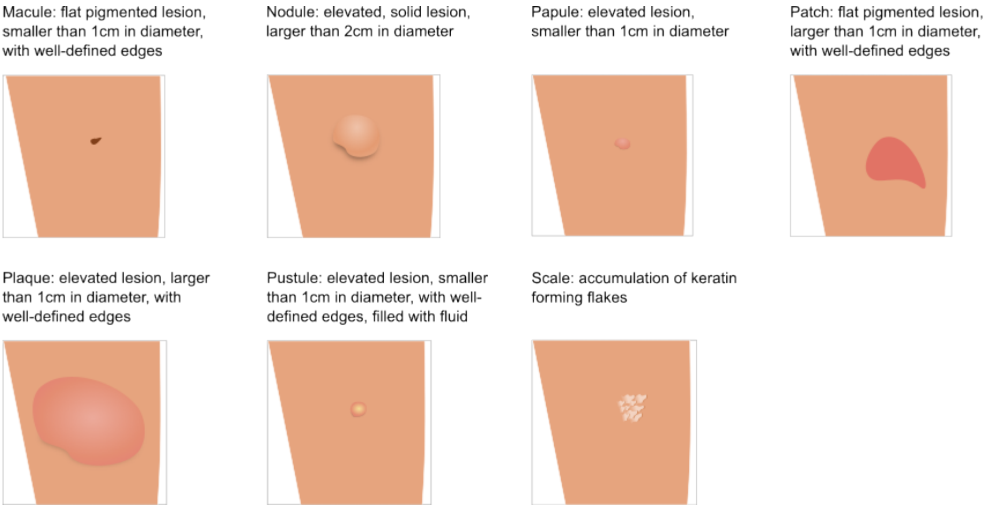

Macule: flat pigmented lesion, smaller than 1cm in diameter, with well-defined edges
Nodule: elevated, solid lesion, larger than 2cm in diameter
Papule: elevated lesion, smaller than 1cm in diameter
Patch: flat pigmented lesion, larger than 1cm in diameter, with well-defined edges
Plaque: elevated lesion, larger than 1cm in diameter, with well-defined edges
Pustule: elevated lesion, smaller than 1cm in diameter, with well-defined edges, filled with fluid
Scale: accumulation of keratin forming flakes

Figure 2.1. Illustration and description of a subset of basic terms for describing skin lesions, as defined by Fitzpatrick and High [2017] and Nast et al. [2016].

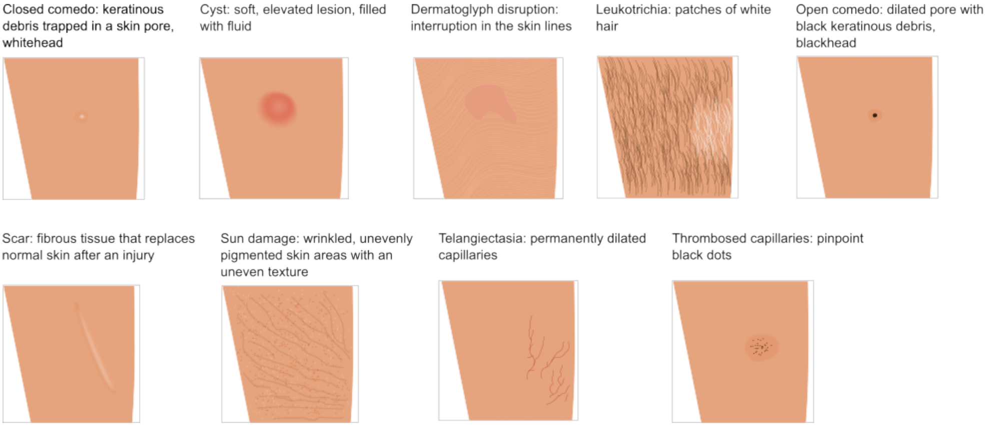

Closed comedo: keratinous debris trapped in a skin pore, whitehead
Cyst: soft, elevated lesion, filled with fluid
Dermatoglyph disruption: interruption in the skin lines
Leukotrichia: patches of white hair
Open comedo: dilated pore with black keratinous debris, blackhead
Scar: fibrous tissue that replaces normal skin after an injury
Sun damage: wrinkled, unevenly pigmented skin areas with an uneven texture
Telangiectasia: permanently dilated capillaries
Thrombosed capillaries: pinpoint black dots

Figure 2.2. Illustration and description of a subset of additional terms for describing skin lesions, as defined by Fitzpatrick and High [2017] and Nast et al. [2016].

General information about a patient can often also help a healthcare professional identify the correct diagnosis. A patient's age, sex, or skin tone can influence a decision due to knowledge about the disease prevalence in specific subpopulations. For example, actinic keratosis is more prevalent in elderly patients, seborrheic dermatitis affects men more often than women, and malignant melanoma is more common in patients with lighter skin tones [Oakley, 2017]. The distribution of lesions over a patient's body can also narrow down the space of possible diagnoses. Chronic plaque psoriasis lesions often appear on extensor sites (elbows, knees), while vitiligo and actinic keratosis lesions are more common in areas exposed to the environment (face, hands).

While a complicated endeavor, diagnosing skin conditions, particularly the ones with specific appearances, can be automated using skin lesion images and ConvNets. In order to take advantage of the recent developments in medical image analysis and the accessibility of smartphones, this thesis focuses on six diseases that can be diagnosed solely from visual markers: acne, actinic keratosis, psoriasis, seborrheic dermatitis, viral warts, and vitiligo [Oakley, 2017].

2.2 Explainability from a human perspective

Since the dawn of time, people have required explanations to satisfy curiosity, examine someone's understanding of a subject, or understand the cause and effect of various phenomena [Miller, 2019]. Explanations are generally given to share knowledge, persuade, or assign blame [Lombrozo, 2006]. Aristotle was the first philosopher to investigate the explanation mechanisms, producing the Four Causes model [Hankinson, 1998] that defines four principal causes that can answer any why question. Since then, many philosophers and researchers have been working on defining when explanations are given, how they are constructed, and how they are communicated.

Although different explanation models have been proposed, the maxims introduced by Grice [1975] lie at the foundation of creating good visual explanations [Miller, 2019]:

- quality: the explanation must be true and well documented;
- quantity: the explanation must include all necessary arguments without including any unnecessary information;
- relation: the explanation must be relevant to the question asked;
- manner: the explanation must be unambiguous, brief, orderly, and only use concepts easily understood by the target of the explanation.

In other words, good explanations must include only clear, high-quality arguments that are relevant to the question, using concepts familiar to the receiver of the explanation. However, it is worth noting that humans are not always able or willing to explain their decisions [Holzinger et al., 2017], which often leads to inconsistent explanations and explanation evaluations by observers.

Given an automated dermatological diagnosis method, we consider its explanation valid if it correctly uses basic and additional terms for describing skin lesions, includes only the relevant descriptors for the chosen diagnosis, and accurately represents the decision mechanism behind the diagnosis. Although mechanisms for producing explanations for an automated method's decision process exist, explanations validated by engineers are often not considered relevant by the target users [Miller, 2019]: domain experts and laypeople tend to expect automated methods to explain their behavior in a way similar to other humans, rather than produce a mathematical description of how the decision process was performed [De Graaf and Malle, 2017]. Figure 2.3 illustrates an explanation for a psoriasis case that fulfills all maxims introduced by Grice when targeted at a healthcare professional.

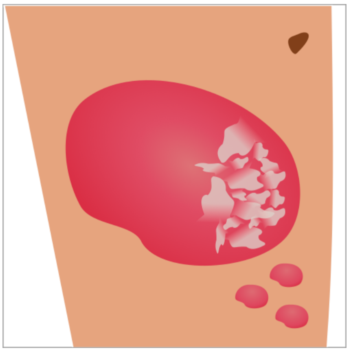


Figure 2.3. Illustration of a psoriasis lesion. A valid explanation for the psoriasis diagnosis would consider the thickened red plaque with well-defined edges, the three thickened red papules with well-defined edges, and the silvery-white scales. The macule at the top right is not indicative of psoriasis and should not be included as an explanation for the diagnosis. A sufficient explanation would only focus on the plaque and the scales, as the two are necessary and sufficient for the diagnosis.

A quantitative evaluation of explanations requires metrics that measure how good an explanation is. In this thesis, we use the metrics introduced by DeYoung et al. [2019] that measure how well an explanation given by an automated method fulfills Grice's maxims:

- plausibility: how similar the given explanation is to that of a human agent;
- faithfulness: how well the explanation represents the inner workings of the automated method.

DeYoung et al. [2019] propose a third metric, sufficiency, that measures whether or not the selected explanation is sufficient for explaining the outcome. We do not include this metric in our work due to how image-level characteristics, such as demographics and body location, can further guide the diagnosis.

As this thesis is concerned with the automated dermatological diagnosis for images of skin diseases, we focus on defining, creating, and evaluating plausible and faithful explanations from the point of view of a dermatologist.

2.3 Explainability approaches in medical imaging

Explainability is an essential factor in the adoption of automated methods in the industry, and this holds even more true within the healthcare system [Kelly et al., 2019]. In the European Union, the General Data Protection Regulation (GDPR) introduced the ”right to an explanation” requirement: any automated system that has an impact on people’s lives must be able to provide high-quality explanations for its behaviour [Goodman and Flaxman, 2017]. This requirement aims to detect and avoid the widespread usage of Clever Hans detectors [Lapuschkin et al., 2019], namely predictors that draw conclusions based on wrong or biased reasoning. One of the most famous examples of a Clever Hans detector is the COMPAS crime recidivism prediction software which was shown to base its decisions on the defendant’s race instead of other, more relevant features [Dressel and Farid, 2018]. Within healthcare, such a detector might learn that cancerous moles are often photographed alongside a ruler and with different pen markings that help with the excision, while non-cancerous moles do not include these elements in their photographs [Winkler et al., 2019]. Instead of focusing on the color and edges of a lesion to assess its risk of malignancy, Clever Hans detectors would focus on whether or not a ruler or any pen markings are present in the image.

While explainability is not an issue in simpler machine learning methods such as linear regression or decision trees, ConvNets are notoriously difficult to explain. This has led to the common adage that there is a trade-off between the machine learning methods’ performance and their explainability: simpler, less performant methods are easily explainable, while larger, more complex methods are more opaque in their decision mechanisms.

Different explanation mechanisms for image analysis deep learning methods have been proposed, offering explanations in visual form, textual form, or through examples [Van der Velden et al., 2022]. Visual explanations are most commonly produced using backpropagation or perturbation. Textual explanations such as image captioning [Vinyals et al., 2015] and example-based explanations such as prototypes [Chen et al., 2019] are used less often in medical research and will not be discussed in this thesis.

Backpropagation-based methods rely on a ConvNet’s backpropagation step to detect the regions in an image with the highest predictive value by computing the gradient in relation to the prediction for all areas of an image. Saliency maps [Simonyan et al., 2013] is a backpropagation-based method that measures how important a pixel is to the prediction by calculating the loss function's gradient for the relevant class with respect to each pixel in the input image. In other words, given an image I of size $m \times n$ and a target class c, we define the saliency map $E_c \in \mathbb{R}^{m\times n}$ for class c as:

$$E_c(I) = \frac{\partial N_c(I)}{\partial I}, \tag{2.1}$$

where $N_c(I)$ is the prediction score for class c of a ConvNet N for the input I.

Gradient class activated maps (Grad-CAM) [Selvaraju et al., 2017] is the most common explainability method in medical imaging [Singh et al., 2020]. It uses the forward propagation step to extract the image features, sets the prediction class to a given value and then backpropagates the signal to the last convolutional layer. The features directly contributing to the predicted class are part of the resulting explanation. More specifically, a Grad-CAM explanation $E_c \in \mathbb{R}^{u\times v}$ for an image I and a given class c is calculated as

$$E_c(I) = \text{ReLU}\left(\sum_k \frac{1}{uv}\sum_{i=1}^{u}\sum_{j=1}^{v}\frac{\partial N_c(A)}{\partial A_{ij}^k}A^k\right), \tag{2.2}$$

where $N_c(A)$ is the class prediction score for c, and $A \in \mathbb{R}^{u\times v\times k}$ is the set of activations for the final convolutional layer in the ConvNet of size $u \times v$ with k filters.

Integrated gradients [Sundararajan et al., 2017] employs a similar mechanism to Grad-CAM. The main difference is that instead of fully calculating the gradient for assigning an importance score to each feature given a convolution layer, the gradient is approximated through a Riemann approximation. Mathematically, an integrated gradients explanation $E \in \mathbb{R}^{u\times v}$ for a set of activations A of size $u \times v$ extracted from an image I using a ConvNet's softmax output $N(I)$ for the input I is expressed as

$$E(I) = \frac{1}{m}(A - A')\sum_{k=1}^{m}\frac{\partial N(I' + \frac{k}{m}(I - I'))}{\partial A}, \tag{2.3}$$

where I' and I' represent the empty image baseline and its final convolutional layer activations, respectively, k is the scaled feature perturbation constant, and m is the number of steps performed during the Riemann approximation of the integral.

As both Grad-CAM and integrated gradients use features extracted from the last convolution layer, their explanations will have a resolution equal to the size of the last layer's features. For example, in an EfficientNet-B2 the resolution of a Grad-CAM explanation is 9×9, while for a VGG the resolution will be 14×14. This results in much less specific explanations than saliency maps, which are computed at the input pixel level.

Perturbation-based methods modify the original image and evaluate the changes in the ConvNet's output with regard to the modifications. The modifications often consist of replacing specific pixels in the image with a constant value. Occlusion [Zeiler and Fergus, 2014] is a straightforward perturbation method that determines which areas of the input image are most relevant for the prediction by systematically deleting sections of the image and computing the impact this alteration has on the prediction confidence. The lower the confidence in the original predicted class, the more important an area is considered to be. Formally, given an image I, a ConvNet N, a class prediction c, and a region R, we define the importance attributed to the region I_R as

$$E_c^R(I) = N_c(I) - N_c(I - I_R), \tag{2.4}$$

where $I - I_R$ is the image I where all pixels in the region I_R have been blocked out, and $N_c(I)$ is the confidence of N when predicting class c for the input I.

Local interpretable model-agnostic explanations (LIME) [Ribeiro et al., 2016] is another popular perturbation method. LIME uses a superpixel algorithm to split the image into areas and then randomly selects a subset of superpixels to occlude. The ConvNet then evaluates the occluded image, and the magnitude of change in its predicted class represents how important the occluded areas are to the prediction. An interpretable model is then trained on a group of perturbed samples and their magnitude of change, resulting in a simplified, explainable model that can be used to better understand the initial ConvNet. Mathematically, LIME generates an explanation E for the ConvNet function N and image I using the explainable model g as

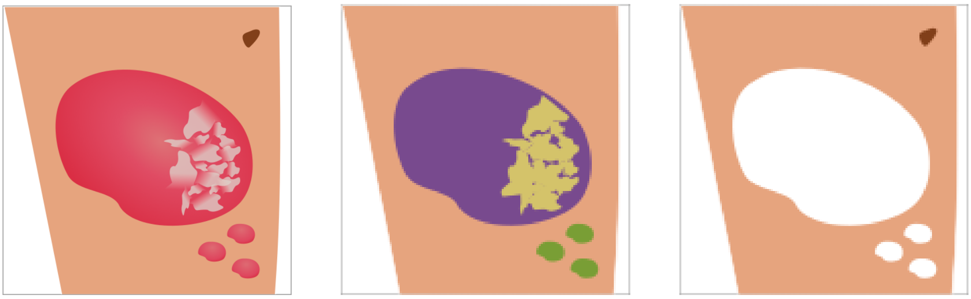
$$E(I) = \arg\min_{g \in G} L(N, g, \pi_I) + \Omega(g), \tag{2.5}$$

where L is the loss function minimized by g, $\Omega(g)$ is the complexity of g, and π_I defines how many perturbations will be explored when training g.

However, in order to obtain stable results, perturbation methods require a large number of inference steps to be performed and thus have high computational requirements. These methods have also been criticized as unsuitable for medical imaging, as the modifications introduce unnatural elements into the image [Uzunova et al., 2019].

Rather than relying on external explainability methods, several groups proposed intrinsically explainable ConvNet architectures, primarily focusing on integrating domain knowledge into the architecture itself. While less transferable than more general techniques, intrinsically explainable ConvNets tend to provide more domain-relevant explanations, often using a vocabulary easily understood by domain experts. Barata et al. [2021] propose an architecture that takes advantage of spatial attention layers and the hierarchical structure of dermatological diagnosis taxonomies to produce more explainable diagnoses. Lin et al. [2022] use hierarchical concept bottleneck models to predict the quality of fetal ultrasound scans by imitating the decision processes steps of radiologists. Gautam et al. [2022] introduce ProtoVAE, a ConvNet architecture that proposes class-specific prototypes and integrates them into the decision process.

Despite the large number of proposed explainability methods, there has been limited research into quantitatively evaluating their explanations. Qualitative analyses rely on visual inspection, often performed by engineers rather than domain experts, while quantitative analyses often use lesion segmentations as the reference standard [Tschandl et al., 2020]. However, such segmentations do not always constitute an adequate explanation. Tschandl et al. [2020] highlight the importance of using the correct reference standard in quantitative analyses through their actinic keratosis example: a Grad-CAM explanation of an actinic keratosis diagnosis highlighted the area surrounding the lesion rather than the lesion itself. While from an engineer's perspective, this behavior suggested that there were issues with the model, dermatologists understood it as the ConvNet focusing on the sun-damaged skin surrounding the lesion, one of the primary characteristics of actinic keratosis, and thus confirmed its correctness.



(a) Original image **(b)** Plausible explanation **(c)** Faithfulness counterfactual

Figure 2.4. Illustrations of plausible, faithful, and sufficient explanations. (a) Original psoriasis image. (b) Ideal plausible explanation. Identical to how a dermatologist would argue for their psoriasis diagnosis, a model offering this explanation uses plaque (purple), scales (yellow), and papules (green) as the reasons for the diagnosis. (c) Counterfactual image for testing the faithfulness of the ideal plausible explanation, created by occluding all characteristics selected by the plausible explanation. The faithfulness test consists of predicting the diagnosis of the counterfactual image and measuring the magnitude of change in the predicted class. If the confidence in psoriasis drops, the explanation is a faithful representation of the ConvNet's decision process.

To evaluate the performance of our explanations, we employ two metrics introduced by DeYoung et al. [2020]: plausibility and faithfulness. We disregard sufficiency due to its limited applicability in dermatology. For quantifying plausibility, we use the fuzzy definitions of sensitivity, specificity, and F1-score (Dice-Sørensen coefficient):

$$\text{F1-score} = \frac{2\sum_{p\in P}\min(E_p, D_p)}{\sum_{p\in P}(E_p) + \sum_{p\in P}(D_p)}, \tag{2.6}$$

$$\text{Sensitivity} = \frac{\sum_{p\in P}\min(E_p, D_p)}{\sum_{p\in P}(D_p)}, \tag{2.7}$$

$$\text{Specificity} = \frac{\sum_{p\in P}\min(1-E_p, 1-D_p)}{\sum_{p\in P}(1-D_p)}, \tag{2.8}$$

where P represents the set of pixels in the image, E defines the set of values for each pixel as given by the explainability method, and D represents the fuzzy dermatologist attention maps calculated as the fraction of dermatologists that included the pixels in their explanation. Similar to occlusion methods, faithfulness can be computed as the magnitude of the change between the predicted class confidence for the original image versus the same class confidence for the image with the explanation areas occluded. Figure 2.4 illustrates the two metrics in a psoriasis case.

Due to differing approaches to generating explanations, methods may produce different explanations for the same ConvNet classification task. In the next section, we compare five explainability methods to evaluate the impact that choosing a certain method would have on the outcome.

2.4 Explainability methods comparison

Different explainability methods take different approaches to explaining a ConvNet's classification. In this section, we explore the differences between a selection of backpropagation-based methods and perturbation-based methods for a six-class diagnosis ConvNet trained on a skin disease dataset achieving 73% test accuracy. We evaluate the explanations by comparing them to explanation maps created by board-certified dermatologists, using the metrics introduced in Section 2.3.

Figure 2.5 illustrates how five common explainability methods approach the explanation of psoriasis diagnosis. In this example, Grad-CAM produces the closest explanations to the dermatologist attention map, closely followed by integrated gradients. Saliency maps and LIME severely underperform. The explanations produced by integrated gradients and Grad-CAM cover large areas of the image due to the low resolution of the last convolutional layer – only 9×9. Saliency focuses on a few highly indicative regions while disregarding other lesions. Both occlusion and LIME tend to focus on irrelevant areas more often, mainly due to the large number of lesions covering a wide area: occluding one lesion is unlikely to impact the diagnosis, as the ConvNet can use the remaining lesions to diagnose the image as psoriasis.

To quantify the overall performance of each explainability method, we produce explanations for 41 correctly classified test images. The average fuzzy sensitivity, specificity, and F1-score obtained by each method are presented in Table 2.1, while the ROC curves for each method and individual dermatologists are illustrated in Figure 2.6. Overall, the same trends emerge: Grad-CAM displays the overall highest sensitivity, F1-score, and ROC AUC, reaching expert-level sensitivity and F1-score. Integrated gradients performs similarly to Grad-CAM. Despite their overall low performance, the high variance allows occlusion and LIME to reach expert-level sensitivity scores. Saliency is the only method that reaches expert-level specificity.

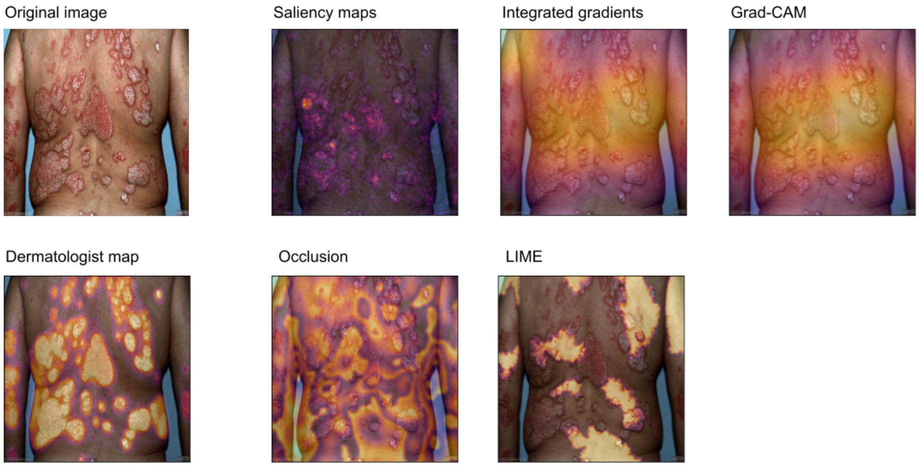


Figure 2.5. Comparison of five explainability methods for the correct psoriasis prediction in a six-class diagnosis ConvNet. Backpropagation-based methods (top row) either pick out single, highly indicative lesions or highlight most plaques, papules, and scales in the image. Occlusion selects both healthy skin and lesions as important for the diagnosis. For occlusion, the large number of lesions and the wide affected area makes it difficult for a single occluded area to impact the diagnosis. LIME focuses mostly on healthy skin, possibly due to its reliance on superpixel algorithms which underperform in images with lower contrast and softer edges.

Table 2.1. Fuzzy sensitivity, specificity, and F1-score for five explainability methods over 41 correctly classified test images. We use the fuzzy dermatologist attention map as a reference standard. The pairwise inter-rater dermatologist agreement is calculated as the average between each dermatologist's explanation with regard to the fuzzy dermatologist explanation map of each other dermatologist. Grad-CAM and integrated gradients reach expert-level sensitivity, although neither can achieve the same results for F1-score due to their low specificity. Saliency obtains a dermatologist-level specificity score. The somewhat low dermatologist sensitivity and F1-score highlight the difficulty of this task.

Method	Sensitivity	Specificity	F1-score
Saliency	0.12 ± 0.05	0.97 ± 0.02	0.17 ± 0.06
Integrated gradients	0.76 ± 0.20	0.69 ± 0.11	0.49 ± 0.19
Grad-CAM	0.81 ± 0.14	0.64 ± 0.12	0.50 ± 0.19
Occlusion	0.69 ± 0.18	0.64 ± 0.10	0.44 ± 0.19
LIME	0.53 ± 0.26	0.45 ± 0.24	0.29 ± 0.17
Dermatologists	0.72 ± 0.04	0.92 ± 0.02	0.67 ± 0.03

These experiments suggest that backpropagation-based methods are more similar to how humans explain their reasoning – either by selecting a few highly indicative

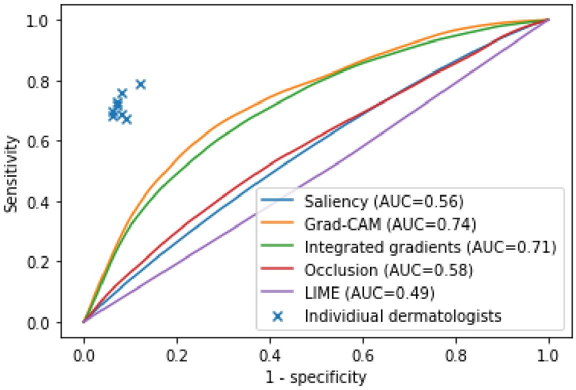


Figure 2.6. ROC curves for five explainability methods over 41 correctly classified test images, and the individual dermatologist pairwise performance with regard to each other dermatologist. Grad-CAM and integrated gradients display the highest performance, while LIME severely underperforms. Dermatologist shows a much higher specificity than all explainability methods, as well as an overall high sensitivity.

regions or by highlighting most of the affected area. Grad-CAM's ease of computation and high performance reinforce the general preference for it in the medical literature. On the other hand, perturbation-based methods seem to not be a good fit for medical applications where a large number of indicative regions may appear in an image.

Another interesting finding was that dermatologists themselves tend to disagree on what a good explanation for a diagnosis is, even when agreeing on the diagnosis. This can be explained by their training, seniority, personality, and even levels of energy when constructing the explanation. Training can influence the vocabulary doctors use in explanations, e.g. some European countries use the term placard (an elevated lesion between 1 and 5cm with well-defined edges) although the term is absent in most North American textbooks. Seniority may enable some dermatologists to easily identify more obscure disease characteristics, while their personality may also influence the explanation: some dermatologists choose to focus on a few, highly indicative lesions, while others tend to select all relevant characteristics present in an image. Finally, as observed in other domains as well, a labeler's ability to focus at the time of performing the task influences the accuracy of their performance – tired, pressured labelers are more likely to make mistakes [Rädsch et al., 2023]. Thus, the creation of a thorough, unified labeling protocol is paramount when attempting to build explainability datasets.

Based on the experiments performed in this section, its low computation requirements [Van der Velden et al., 2022], and its popularity within the field of medical imaging [Singh et al., 2020], we use Grad-CAM as the main explainability method in this thesis.

2.5 Convolutional neural networks in dermatology

Given the large proportion of people experiencing skin diseases, the low number of dermatologists per capita, and the availability of imaging devices, skin lesion diagnosis is one of the prime targets for automation within healthcare. In 2017, Esteva et al. [2017] proved that a ConvNet could achieve a diagnosis performance similar to that of dermatologists, thus paving the way toward the current explosion in ConvNets applied to dermatological tasks. This development is in part supported by the release of large skin disease imaging datasets such as SD-260 [Sun et al., 2016], the International Skin Imaging Collaboration (ISIC) challenge dataset [Codella et al., 2018], and DermNetNZ [DermNetNZ, 2021].

Esteva et al. [2017] triggered an explosion in the study of automated methods within dermatology. A large variety of modalities, tasks, and architectures have been tackled by researchers in machine learning applications for dermatology [Thomsen et al., 2020]. Dermoscopic images were one of the first modalities to be investigated by the machine learning community. They are images taken with a dermoscope by a healthcare professional, often in a clinical setting. Dermoscopic images are primarily used in diagnosing different types of skin cancer and their mimickers, e.g. actinic keratosis, seborrheic keratosis, benign naevi [Esteva et al., 2017, Tschandl et al., 2020, Reshma et al., 2022]. With the release of SD-260 and DermNetNZ, photographic images taken with professional cameras in a clinical environment have also become a staple of dermatological automation research. Taken in standardized poses and controlled lighting and background settings, they have wider use cases than dermoscopy imaging, often focusing on the diagnosis of chronic skin diseases, severity assessments, or progression analyses [Zhang et al., 2019, Han et al., 2020, Wu et al., 2021, Aggarwal and Papay, 2022, Ba et al., 2022]. As smartphones became more ubiquitous, photographic images taken by patients themselves using a smartphone made their entrance into the research community [Jensen et al., 2019, Chin et al., 2020, Hossain et al., 2022]. However, the differences in smartphone camera quality, the lack of standardized poses, and the diverse lighting settings make their use more challenging.

Tasks tackled by ConvNets cover a large variety of dermatological applications. Disease diagnosis either focuses on binary disease classification or distinguishing between different skin diseases [Tschandl et al., 2020, Reshma et al., 2022, Han et al., 2020, Aggarwal and Papay, 2022, Hossain et al., 2022]. In diseases such as acne, lesion count [Min et al., 2013] is a common task, while lesion segmentation [Yuan et al., 2017, Baig et al., 2020, Wu et al., 2022] is relevant in cases such as skin cancer where the lesion size and its edges may further inform the diagnosis. Disease severity assessment [Seité et al., 2019, Munthuli et al., 2022, Zhang and Ma, 2022] is often used for triaging patients or to establish the right treatment. Within teledermatology, ConvNets have been used in image quality assessment [Kim and Lee, 2017, Bianco et al., 2018].

Most solutions use standard ConvNet architectures pre-trained on the ImageNet dataset [Deng et al., 2009], such as VGG [Han et al., 2020, Chin et al., 2020], ResNet50 [Hossain et al., 2022, Burlina et al., 2019], MobileNet [Hossain et al., 2022, Sahin et al., 2022], or EfficientNet [Ba et al., 2022, Hossain et al., 2022, Sahin et al., 2022]. The preference towards pre-training can be explained by the small size of public datasets: a ConvNet would have difficulties being trained from scratch on a dataset with fewer than 30,000 images, and thus pre-training is necessary to ensure stability and performance.

This exploration of dermatological modalities and tasks using ConvNets is not limited to academia – different industrial entities are also investigating ConvNets for dermatology. La Roche Possay [Seité et al., 2019], LEO Pharma [Jensen et al., 2019], Google [Jain et al., 2021], Nurithm Labs [Shah et al., 2021], and L'Oréal [Flament et al., 2022] have all proposed ConvNets for industrial applications. Their work covers topics such as disease severity, diagnosis confidence calibration, diagnosis performance, and skin aging evaluation. However, none of their proposed methods have as of yet been certified as a medical device for public-facing applications.

To further understand the inner working of their proposed ConvNets, some groups include an explainability method to provide additional insight into their proposal's decision mechanism. The most common explainability approaches are based on backpropagation, such as saliency maps [Fink et al., 2020, Liu et al., 2020], Grad-CAM [Tschandl et al., 2020, Zunair and Hamza, 2020, Xie et al., 2020, Tanaka et al., 2021], and trainable attention [Barata et al., 2021, Yan et al., 2019].

In this thesis, we explore the explainability of pre-trained ConvNet architectures and employ custom-built architectures to obtain highly performant, explainable skin lesion diagnosis methods.

CHAPTER 3

An end-to-end methodology for explainable convolutional neural networks

Based on work done by Raluca Jalaboi, Frederik Faye, Mauricio Orbes-Arteaga, Dan Jørgensen, Ole Winther, and Alfiia Galimzianova. Published in the Medical Image Analysis Journal, 2023. See Appendix A.

As highlighted in Chapter 2, explainability is one of the main obstacles towards the adoption of automated methods in the healthcare system. In this chapter, we explore explainability from the point of view of a dermatologist by creating a skin disease diagnosis explainability dataset (Section 3.1) and afterward use this dataset to train two intrinsically explainable ConvNet architectures (Section 3.2).

3.1 Explainability dataset for skin disease diagnosis

Despite the large variety of explainability methods proposed to demystify the inner workings of ConvNets, little focus has been given to understanding how experts explain their diagnoses and how their explanations relate to those of automated methods. To this end, we created DermXDB: a dermatological diagnosis explainability dataset annotated by eight board-certified dermatologists. After several steps of onboarding to ensure that we achieve the highest possible label quality, dermatologists were asked to diagnose 554 images sourced from public skin disease datasets [Sun et al., 2016, DermNetNZ, 2021] and annotate them with explanations in the form of skin lesion characteristics [Nast et al., 2016]. 65% of the images depict patients with light skin, 32% depict patients with a medium skin tone, and dark skin patients represent 3% of the dataset. 19% of the patients were young, 55% adults, and 26% elderly.

The dataset includes six disease classes: acne, actinic keratosis, psoriasis, seborrheic dermatitis, viral warts, and vitiligo. Skin lesion characteristics were selected from the literature [Oakley, 2017] and were reviewed by two board-certified dermatologists with more than ten years of experience. This step ensured that the resulting explanations used a common vocabulary with domain experts. Figure 3.1 shows the explanation taxonomy for localizable characteristics, while Figure 3.2 and Figure 3.3 show image-level characteristics and additional descriptive terms for the localizable characteristics taxonomies, respectively. The dense annotation protocol allowed us to create a thorough, structured database of explanations for the six selected diseases.

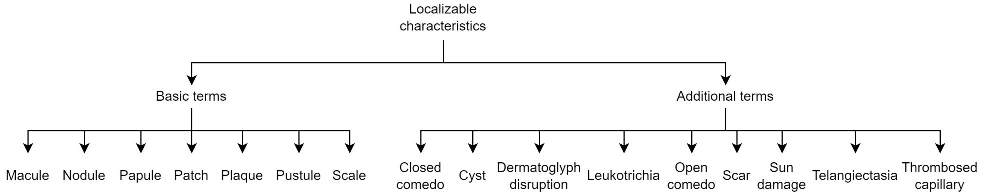


Figure 3.1. Localizable characteristics taxonomy.

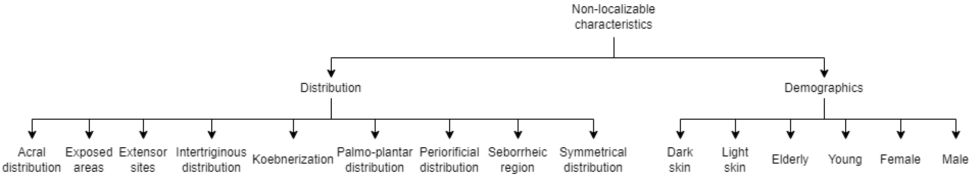


Figure 3.2. Image-level characteristics taxonomy.

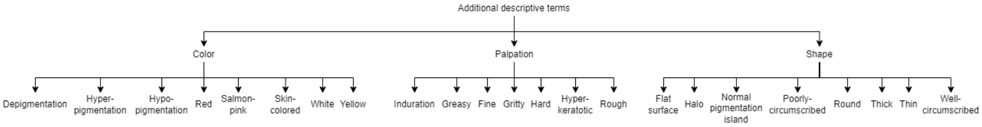


Figure 3.3. Taxonomy for the additional descriptive terms associated with the localizable characteristics.

The difficulties surrounding the generation of explanations described in Chapter 2 were observed when comparing the explanations given by dermatologists for the same image: dermatologists tend to disagree on how to explain a diagnosis, even when they agree on the diagnosis itself. This suggests that in order to accurately evaluate an explainability method's performance, we need to consider the opinions of multiple domain experts.

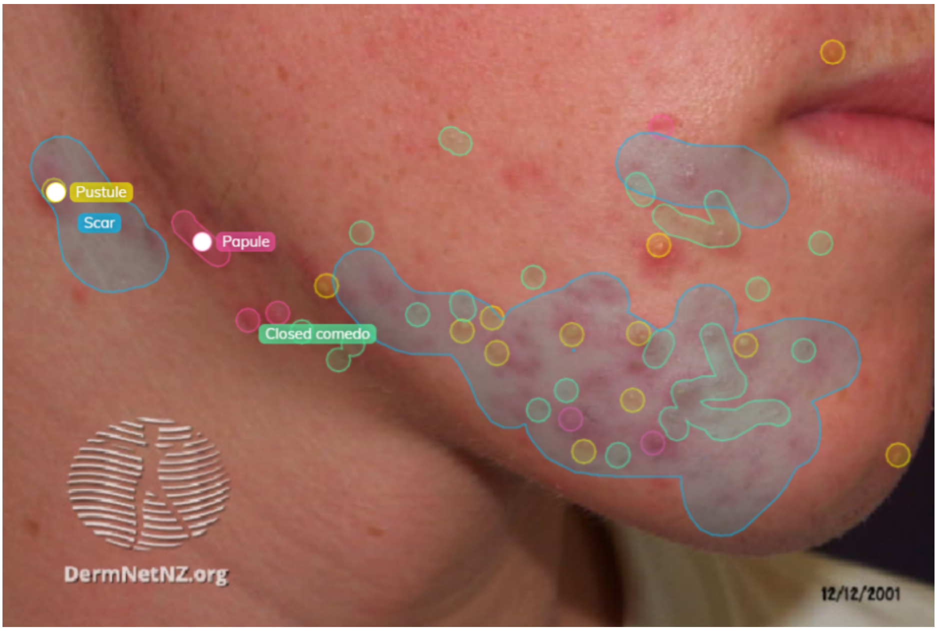


Figure 3.4. Characteristics labeled by a dermatologist for an acne case. Following the instructions, no characteristic was segmented, but rather the region where they were present was identified without necessarily following the lesion boundaries. For more difficult characteristics to locate, e.g. scars, dermatologists were instructed to brush over entire areas containing the characteristic. The labeling interface was provided by V7-Labs [2021].

All explanations created as part of this dataset are complete rather than just sufficient due to differences in explanation needs for users from different backgrounds [Van der Velden et al., 2022]. A dermatologist might prefer a sufficient explanation for a psoriasis diagnosis (e.g. the presence of a red, indurated plaque with well-defined edges and thick, silvery-white scales), while a general practitioner or a patient might prefer a complete explanation (e.g. the presence of multiple red, indurated plaques and papules with well-defined edges, thick, silvery-white scales, and hypopigmented patches). Having a dataset that consists of complete explanations enables us to pick the correct explanation for the target audience of our diagnosis ConvNets. Figure 3.4 illustrates one dermatologist's explanation process for an acne case. Note the adherence to Grice's maxims:

- quality: the explanation is created by a board-certified dermatologist for the reference standard diagnosis of acne;

- quantity: all necessary characteristics for this diagnosis were marked, while irrelevant characteristics (e.g. freckles) were disregarded;
- relation: the explanation only concerns the acne diagnosis;
- manner: the explanation uses dermatological concepts derived from the medical literature.

With DermXDB, we introduced the first publicly available dermatological diagnosis explainability dataset, created in collaboration with domain experts. Its release enabled the creation of intrinsically explainable ConvNet architectures, as shown in Appendix A, and the benchmarking of classical explainability methods, as presented in Appendix B.

3.2 DermX: an intrinsically explainable architecture

With the help of the explainability dataset introduced in Section 3.1, we developed DermX, an intrinsically explainable skin lesion diagnosis architecture. DermX+ expanded upon DermX by adding a guided attention component to enhance the characteristic localization.

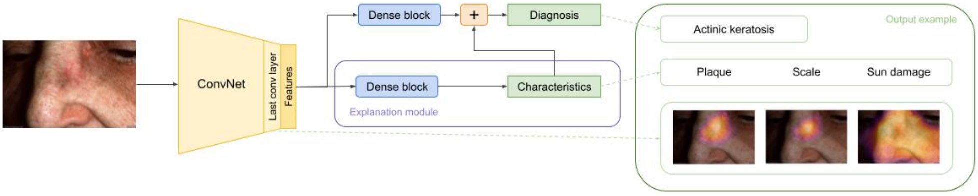


Figure 3.5. Clinically-inspired convolutional neural network architecture for image diagnosis with explanations in the form of skin lesion characteristics. Given an image, the model is trained to predict the diagnosis together with the supporting characteristics. The diagnosis is predicted using the characteristics identified by the model (similar to how dermatologists diagnose cases) and the extracted image features. Using the extracted features alongside the predicted characteristics ensures no relevant information is lost, e.g. the age or the skin tone.

Figure 3.5 and Figure 3.6 illustrate the DermX and DermX+ architectures, respectively, and highlight the explanation module of each ConvNet. Our clinically-inspired ConvNets take advantage of the explainability labels introduced by DermXDB (Section 3.1) and enable emulating the explanation mechanisms of a dermatologist. We evaluated both architectures for diagnosis performance, explanation plausibility, and explanation faithfulness, as described in Chapter 2.

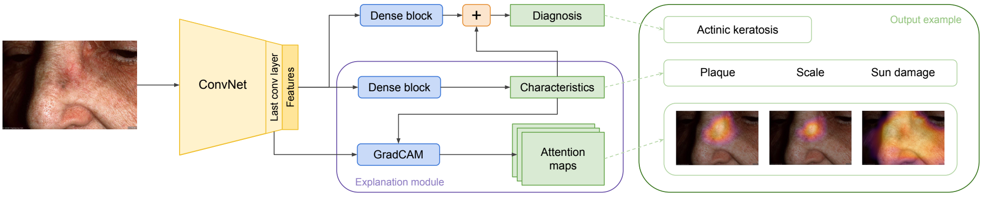


Figure 3.6. Explainable convolutional neural networks architecture for skin disease diagnosis. DermX+ expands upon the DermX architecture presented in Figure 3.5 by introducing a guided attention components: the network now learns where each relevant characteristic is located in addition to the diagnosis and its supporting characteristics.

Our results showed that both DermX and DermX+ obtained an almost expert-level explainability performance without sacrificing their diagnosis performance. Figure 3.7 illustrates the differences between DermX and DermX+ on characteristic localization: DermX included slightly more irrelevant information, while DermX+ closely followed the dermatologist outlines. The faithfulness analysis proved that both DermX and DermX+ used the predicted characteristics to decide on the diagnosis and thus that the characteristics and their localizations were accurate explanations for the decisions taken.

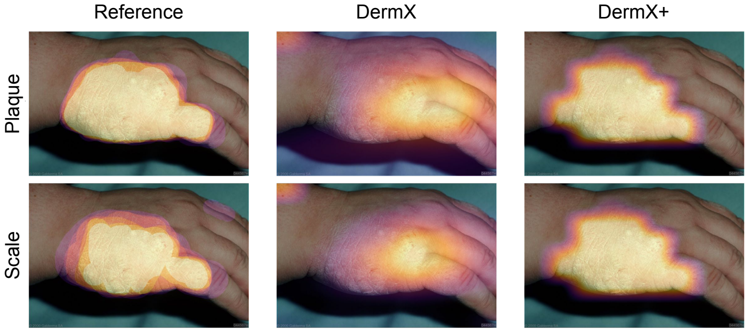


Figure 3.7. Characteristic attention maps for a correctly classified psoriasis case for the two identified characteristics: plaque (first row) and scale (second row). The first column shows the dermatologist-derived fuzzy attention map, the second one illustrates the Grad-CAM for each characteristic generated by DermX, while the last column shows the DermX+ Grad-CAM maps. DermX+ displays much closer results to the reference standard, while DermX maps include more irrelevant information, such as finger knuckles.

The two proposed architectures, particularly DermX+, follow Grice's maxims by providing faithful explanations that include most arguments deemed relevant by a dermatologist. All concepts used in the explanations are intrinsically understandable to dermatologists through how the dataset was created.

With DermX and DermX+, we proved that we can develop intrinsically explainable ConvNets that reach expert-level explanation performance without sacrificing the classification performance.

CHAPTER 4

Explainability benchmark

Based on work done by Raluca Jalaboi, Ole Winther, and Alfiia Galimzianova. See Appendix B.

With the release of DermXDB (described in Chapter 3.1), a quantitative analysis of disease diagnosis ConvNet explainability became possible. To evaluate how explainable ConvNet architectures are, we identified the open-access architectures used in the literature for skin disease diagnosis on natural photography, and generated explanations for their diagnoses using Grad-CAM.

We selected the benchmarked architectures through a systematic literature review, following the methodology introduced by Thomsen et al. [2020]. 22 articles were included in the analysis, covering 11 ConvNet architectures. Our literature review revealed the impact of Esteva et al. [2017] in medical imaging: the number of manuscripts focusing on dermatological applications skyrocketed after 2017. Additionally, we noticed an increase in industrial involvement in research starting in 2019 and an exploration of different modalities and tasks beginning in the same year (see Figure 4.1). Despite these developments, no automated methods for dermatological tasks are currently available to the public. These findings reinforce the importance and value of automation in the healthcare system and its slow acceptance within the field.

To create the benchmark, all 11 architectures were pre-trained on a proprietary clinical skin disease dataset and fine-tuned on a subset of DermXDB, which allowed for higher performance in all ConvNets by reducing the domain shift. We evaluated each ConvNet's explainability by extracting Grad-CAM attention maps and comparing them to the DermXDB dermatologist attention maps. In other words, we evaluated the plausibility of the Grad-CAM explanations by comparing them with a domain expert's explanations. No ConvNet achieved expert-level diagnosis performance on the overall dataset, although most did achieve expert-level performance when diagnosing actinic keratosis and seborrheic dermatitis. The explainability results are similar, although some ConvNets outperformed dermatologists at a characteristic level. Figure 4.2 illustrates the relationship between the ConvNet diagnosis performance and their explainability F1-scores.

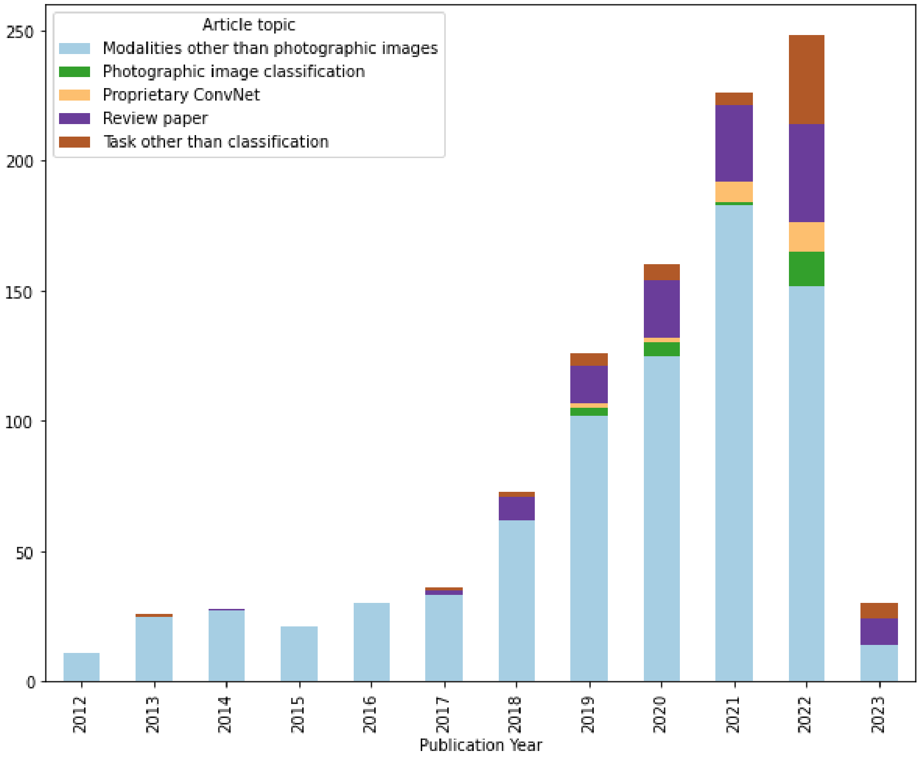


Figure 4.1. Distribution of retrieved article topics per publication year (search query ran on the 20th of February 2023). 2018 marks an explosion in the number of deep learning applications in dermatology, a fact highlighted by the large increase in articles in the subsequent years and an increase in review articles. In 2019, the industrial involvement in this field first became apparent with an increase in proprietary ConvNets. 2019 also marks the first emergence of dermatological applications using photographic imaging. Finally, although classification is still the most common application, other applications are becoming increasingly more researched.

The differences between what different ConvNets focus on are illustrated in Figure 4.3: more modern architectures focused on the entire affected area, while older architectures focused on single, highly indicative lesions. In some cases, we observe a tendency to follow a Clever Hans detector approach to classification: some ConvNets focus on the lips when diagnosing acne (as acne is most often encountered on a patient's face) or the watermark for vitiligo (where most samples were extracted from the DermNetNZ dataset). This evaluation of explanations lends further weight to GDPR's right to an explanation by highlighting the situations where a ConvNet might produce a diagnosis using irrelevant arguments. Additionally, our results reinforce the importance of using explainability methods such as Grad-CAM to better understand the inner

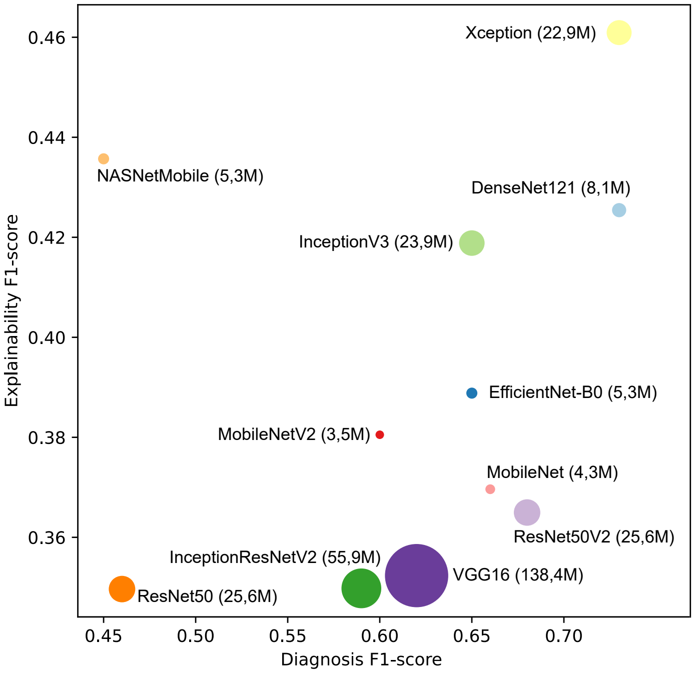


Figure 4.2. ConvNet explainability as a function of ConvNet performance and the number of parameters. Xception displays both the highest performance and image-level explainability, while ResNet50 performs poorly in both criteria.

workings of a ConvNet's decision mechanisms.

Our findings confirm the observations described in Chapter 3.1 surrounding the difficulty in proposing explanations: both dermatologists and ConvNets take different approaches to their explanations, and thorough quantitative assessments are necessary to accurately evaluate how well their explanations fulfill Grice's maxims. These results highlight that even though ConvNets can produce plausible explanations, more work is needed to achieve expert-level performance. Additionally, more focus should be given to developing intrinsically explainable ConvNet architectures rather than relying on post hoc explainability methods. When comparing the benchmarked architectures to the two ConvNets proposed in Chapter 3.2, we find that DermX obtains a slightly higher sensitivity for characteristic localization than the best-performing benchmarked architecture, while DermX+ shows a much lower sensitivity. This can be explained through the high specificity of DermX+, which tends to follow the dermatologist-derived attention maps more closely than other architectures.

Overall, this work accentuates the need for further research into the field of explainability: similar datasets in different domains are needed to allow researchers and engineers to pick the right architecture for the tasks at hand.

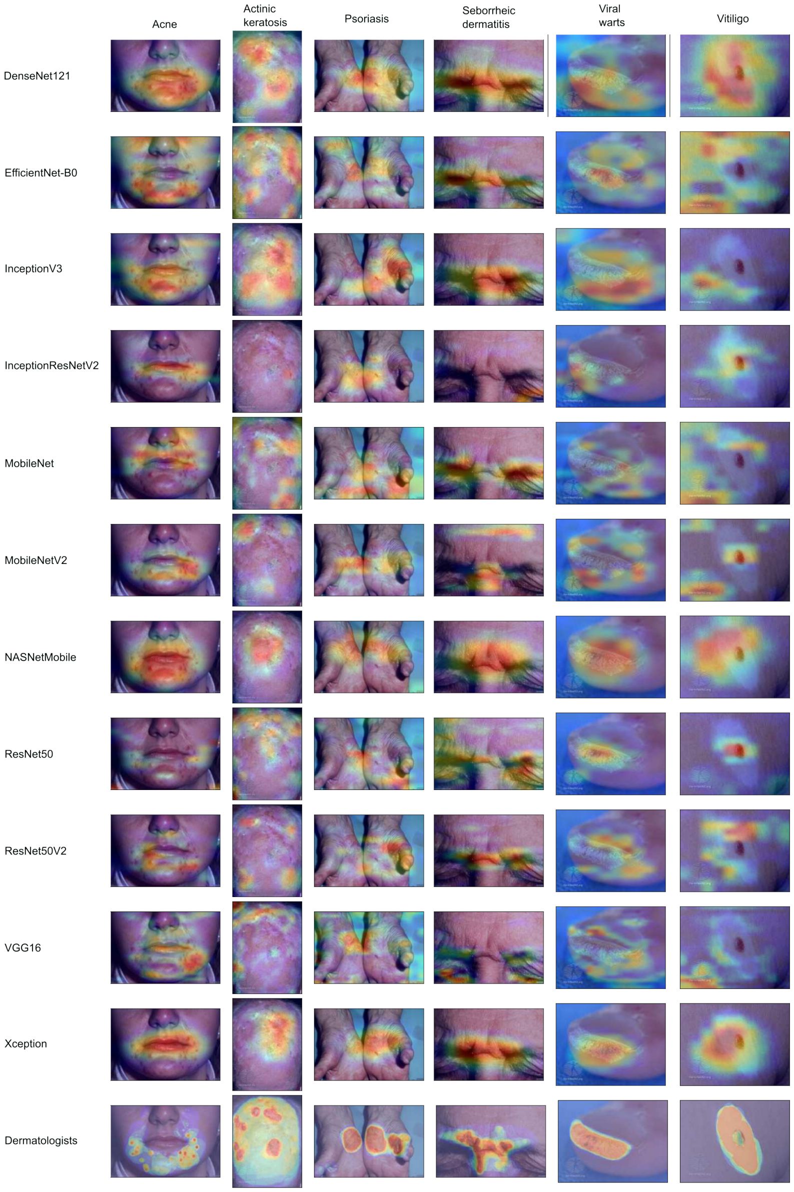


Figure 4.3. Example of Grad-CAM outputs for six images correctly diagnosed by all ConvNets. Older ConvNets, such as VGG16, ResNet50, ResNet50V2, and InceptionResNetV2, tend to focus on a single, highly indicative lesion rather than the whole affected region. More modern ConvNets, such as NASNetMobile, Xception, and EfficientNet, focus on the entire affected area. Some ConvNets overfitted during training and focus on the watermark when diagnosing vitiligo.

CHAPTER 5

Applying DermX in teledermatology

Based on work done by Raluca Jalaboi, Ole Winther, and Alfiia Galimzianova. Published in the Telemedicine and e-Health Journal, 2023. See Appendix C.

Within teledermatology, the need for high-quality images is paramount. However, up to 30% of the images sent by patients are low-quality, requiring retaking before dermatologists can diagnose them [Pasquali et al., 2020]. One way to achieve better image quality is by deploying on-device explainable image quality assessment methods. To this end, we applied the methodology described in Chapter 3 to create a ConvNet architecture for explainable image quality assessment.

We used a dataset of smartphone images taken by patients and labeled by up to 12 board-certified dermatologists with either a diagnosis, a rejection class (no skin visible, no lesion visible), or at least one image quality issue (bad framing, bad light, blurry, low resolution, too far away). Figure 5.1 illustrates the five image quality issues included in the dataset. The data collection protocol followed the one introduced in Chapter 3.1. With this dataset, we trained ImageQX, an explainable image-quality ConvNet architecture inspired by DermX.

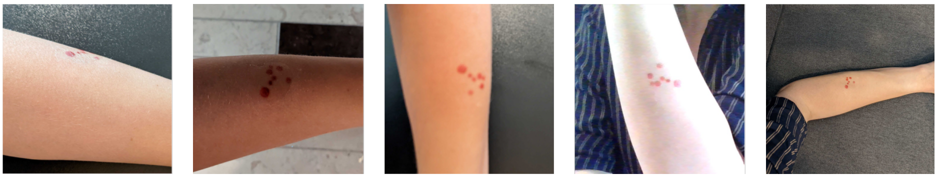


Figure 5.1. Illustration of poor image quality explanations that ImageQX can detect. (a) *Bad framing*: the image was not centered on the lesion. (b) *Bad light*: the lighting conditions in which the image was taken were too dark. (c) *Blurry*: the image is not focused on the lesion, masking out its details. (d) *Low resolution*: the image was taken with a low-resolution camera, and few details can be discerned. (e) *Too far away*: few lesion details could be seen due to the distance from the camera. Images courtesy of the authors.

Our results were similar to the ones described in Chapter 3.2: we obtain expert-level explainability performance at no cost to the classification performance. Using Grad-CAM, we were also able to locate the area of an image that influenced the prediction of a quality issue. Figure 5.2 illustrates this localization process on a blurry image.

The high classification performance allowed around 70% of low-quality images to be identified before reaching the dermatologists. With a size of only 15MB, ImageQX is easily deployable directly on mobile devices, enabling instant feedback regarding the image quality. Access to additional information about what image quality issues were identified in the image and their location could offer patients guidance on improving their image quality and thus reduce the time to diagnosis and treatment.

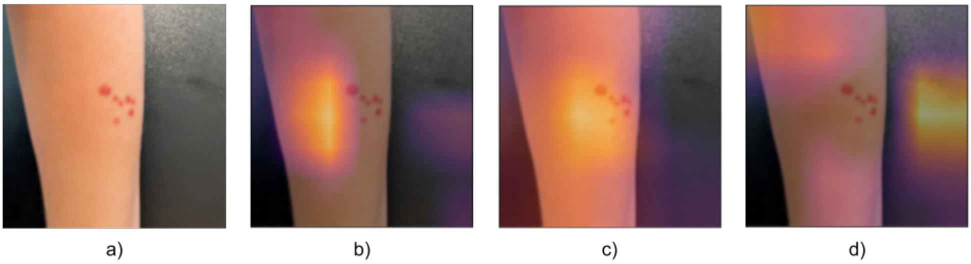


Figure 5.2. Grad-CAM attention maps for the blurry test image introduced in Figure 5.1. The image was correctly classified as poor quality. (a) the original blurry image. (b) Grad-CAM attention map for *bad light*. (c) Grad-CAM attention map for *blurry*. (d) Grad-CAM attention map for *low resolution*. When predicting *bad light*, ImageQX focuses on a slightly shaded part of the arm, while for *blurry* it highlights the lesion and its surrounding area. The *low resolution* prediction is based on the edges of the arm and the background. Image courtesy of the authors.

Through this work, we demonstrated that the methodology introduced by DermX can be applied to other domains with similar performance. Implementing ImageQX in the teledermatology flow could reduce the burden on dermatologists by removing the majority of low-quality images from their flow while at the same time reducing the time to a patient being diagnosed and treated.

CHAPTER 6

Conclusion

In this chapter, we summarize the contributions presented in this thesis, discuss the results and their impact on society, and suggest possible avenues for future work.

6.1 Contributions

The goal of this thesis was to investigate explainable ConvNets for skin lesions diagnosis. To this end, we developed an end-to-end methodology for explainable ConvNets (Chapter 3). We started by building DermXDB, a skin lesion diagnosis explainability dataset (Chapter 3.1). DermXDB uses terminology derived from the medical literature and was annotated by eight board-certified dermatologists, enabling us to measure the expert-level agreement on diagnosis explanations. We found that even when dermatologists agree on a diagnosis, their explanations might differ, which suggests that a thorough evaluation of an explainability method must consider the opinions of multiple domain experts. Afterward, we proposed DermX and DermX+, two intrinsically explainable ConvNet architectures trained on the DermXDB dataset (Chapter 3.2). DermX learned to diagnose images and explain its decision in a way similar to that of dermatologists. DermX+ expanded upon DermX by introducing a guided attention component that also learns the location of the identified characteristics. Our results contradict the common adage that explainability must come at the cost of performance: the two methods achieved similar performance to a classical diagnosis model while providing plausible and faithful explanations.

Using DermXDB, we benchmarked the explainability of ConvNet architectures commonly used in skin lesion diagnosis for photographic images (Chapter 4). We compared Grad-CAM attention maps for each ConvNet to the fused explanation maps created by dermatologists in DermXDB. Despite approaching expert-level performance, no network fully achieves it. This finding highlights the importance of building intrinsically explainable ConvNet architectures rather than solely relying on general explainability techniques.

Finally, we proved the generalizability of the DermX methodology by applying it to a different application – teledermatological image quality assessment (Chapter 5). In this work, we collected smartphone images taken by patients and asked up to 12 board-certified dermatologists to diagnose them in a teledermatology setting or label them with quality issue tags if they were too low quality to diagnose. ImageQX, a ConvNet architecture inspired by DermX, was able to filter more than 70% of the images tagged as low-quality while also achieving expert-level performance in detecting the quality issues present in the image. With a size of only 15MB, ImageQX can easily be deployed on mobile devices and thus be seamlessly integrated within the teledermatology flow, reducing the burden on dermatologists and the time to diagnosis and treatment for patients.

6.2 Discussion

Throughout this thesis, we took a slightly different approach to explainability than other works within the domain. Rather than focusing on creating a new explainability method, we investigated what explainability means to domain experts and how we can produce explanations that are acceptable to them. Our findings show that training the ConvNets to produce plausible explanations can achieve near-expert-level performance in both diagnosis and explanations. While this procedure incurs no diagnosis performance penalty, there is still a cost to be paid: creating explainability datasets is expensive and time-consuming, as it requires input from both machine learning researchers on how to structure the data and from domain experts to define the terminology and annotate the data.

Our work has several industrial applications. Within diagnosis, DermX and DermX+ can serve as trusted second opinions for healthcare professionals due to their high diagnostic performance and plausible explanations. With additional validation and the appropriate certifications, DermX and DermX+ could be candidates for deployment in a patient-facing diagnosis application. ImageQX may be used to improve the image quality in a teledermatology flow by deploying it directly on patient mobile devices and guiding them to take high-quality pictures through personalized feedback.

In future work, we will expand DermXDB by introducing more disease classes and exploring more application domains, e.g. breast cancer scans, Alzheimer's detection in brain scans. Additionally, we would like to use image-level explanations and additional descriptive terms to build ConvNets capable of producing more comprehensive and accurate explanations.

Our research emphasizes the need for more publicly available explainability datasets: they enable the thorough benchmarking of explainability methods and the training of explainable architectures displaying expert-level performance. Such datasets are especially valuable in domains where explainability is required for adoption, such as medicine, justice, or finance. We believe that the methodology we introduced can both serve as a basis for future developments in these areas, as well as enrich the field of explainability in medical imaging.

APPENDIX A

DermX: An end-to-end framework for explainable automated dermatological diagnosis

Published in the Medical Image Analysis Journal, January 2023.

Medical Image Analysis 83 (2023) 102647

Contents lists available at ScienceDirect

Medical Image Analysis

journal homepage: www.elsevier.com/locate/media

DermX: An end-to-end framework for explainable automated dermatological diagnosis

Raluca Jalaboi [a,b,*], Frederik Faye [b], Mauricio Orbes-Arteaga [b], Dan Jørgensen [b], Ole Winther [a,c,d], Alfiia Galimzianova [b]

[a] Department of Applied Mathematics and Computer Science at the Technical University of Denmark, Richard Petersens Plads, Building 324, DK-2800 Kongens Lyngby, Denmark
[b] Omhu A/S, Silkegade 8 st, DK-1113 Copenhagen C, Denmark
[c] Bioinformatics Centre, Department of Biology, University of Copenhagen, Copenhagen, Denmark
[d] Center for Genomic Medicine, Rigshospitalet, Copenhagen University Hospital, Copenhagen, Denmark

ARTICLE INFO

MSC:
92B20
92C50
68T45

Keywords:
Dermatology
Explainability
Dataset
Convolutional neural networks

ABSTRACT

Dermatological diagnosis automation is essential in addressing the high prevalence of skin diseases and critical shortage of dermatologists. Despite approaching expert-level diagnosis performance, convolutional neural network (ConvNet) adoption in clinical practice is impeded by their limited explainability, and by subjective, expensive explainability validations. We introduce DermX, an end-to-end framework for explainable automated dermatological diagnosis. DermX is a clinically-inspired explainable dermatological diagnosis ConvNet, trained using DermXDB, a 554 image dataset annotated by eight dermatologists with diagnoses, supporting explanations, and explanation attention maps. DermX+ extends DermX with guided attention training for explanation attention maps. Both methods achieve near-expert diagnosis performance, with DermX, DermX+, and dermatologist F1 scores of 0.79, 0.79, and 0.87, respectively. We assess the explanation performance in terms of identification and localization by comparing model-selected with dermatologist-selected explanations, and gradient-weighted class-activation maps with dermatologist explanation maps, respectively. DermX obtained an identification F1 score of 0.77, while DermX+ obtained 0.79. The localization F1 score is 0.39 for DermX and 0.35 for DermX+. These results show that explainability does not necessarily come at the expense of predictive power, as our high-performance models provide expert-inspired explanations for their diagnoses without lowering their diagnosis performance.

1. Introduction

Skin diseases affect a third of the global population (Hay et al., 2014) and are the fourth leading cause of disability worldwide (Karimkhani et al., 2017). The increasing demand for dermatological care is exacerbated by the low performance of general practitioners when diagnosing skin conditions (Federman et al., 1999), and by the global scarcity of expert dermatologists (Feng et al., 2018; Kringos et al., 2015).

Automation may help alleviate this problem. Convolutional neural networks (ConvNets) have been shown to achieve near expert-level performance in diagnosing dermatological conditions from images of skin lesions (Thomsen et al., 2020; Esteva et al., 2017), and that they are able to assist general practitioners as well as less experienced dermatologists in improving their diagnostic performance (Tschandl et al., 2020; Jain et al., 2021). However, the lack of a good explanation mechanism (Kelly et al., 2019) for ConvNet decisions is one of the main obstacles to their adoption as automated diagnosis systems (Goodman and Flaxman, 2017; Kelly et al., 2019; Topol, 2019). A good explanation is expected to be both *plausible*, i.e. as similar as possible to a human explanation, and *faithful*, i.e. to accurately represent the inner workings of the network (Jacovi and Goldberg, 2020).

Different mechanisms for explaining ConvNet decisions have been proposed (Simonyan et al., 2014; Selvaraju et al., 2017; Ribeiro et al., 2016). Within the medical imaging literature, the most common explainability methods are saliency-based methods, such as raw saliency maps (Simonyan et al., 2014) and gradient-weighted class-activation attention maps (Grad-CAM) (Singh et al., 2020). While other methods were criticized due to their lack of faithfulness, Grad-CAMs have been shown to perform well (Adebayo et al., 2018). However, there remains a lack of standard metrics for plausibility validation, as the explanations they provide are often incomplete and difficult to quantify (Tschandl

[*] Corresponding author at: Department of Applied Mathematics and Computer Science at the Technical University of Denmark, Richard Petersens Plads, Building 324, DK-2800 Kongens Lyngby, Denmark.
E-mail address: rjal@dtu.dk (R. Jalaboi).

https://doi.org/10.1016/j.media.2022.102647
Received 15 December 2021; Received in revised form 17 August 2022; Accepted 27 September 2022
Available online 7 October 2022

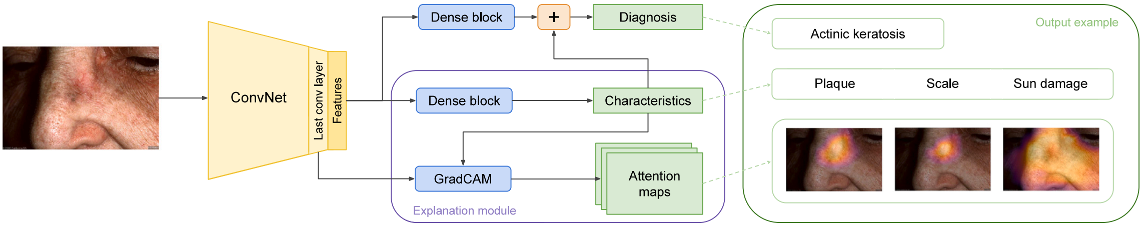


Fig. 1. Clinically-inspired convolutional neural network architecture for image diagnosis with explanations in the form of skin lesion characteristics. Given an image, the model is trained to predict the diagnosis together with the supporting characteristics, and to focus its attention on image sections that contain relevant characteristics. The diagnosis is predicted using both the characteristics identified by the model (similar to how dermatologists diagnose cases), and the extracted image features. Using the extracted features alongside the predicted characteristics ensures that no relevant information is lost, e.g. the age or the skin tone. The explanation module offers plausible, faithful explanations to the diagnosis predicted by the model, while also localizing the explanations in the image.

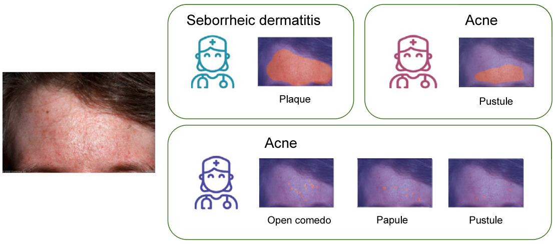


Fig. 2. Sample from the DermXDB dataset. A seborrheic dermatitis image from the SD-260 dataset was evaluated by eight dermatologists. Three evaluations are depicted in this figure. One dermatologist correctly diagnosed it as seborrheic dermatitis due to the presence of plaque. Another dermatologist incorrectly diagnosed it as acne due to the presence of open comedones, papules, and pustules, while a third dermatologist diagnosed it as acne due to the presence of pustules.

et al., 2020). More specifically, common ConvNet explainability methods provide no semantic information alongside the explanation, but rather focus on the image section where the network pays attention. In complex domains such as dermatology, this information is not enough to explain the decision mechanisms: knowing that the network focuses on the skin lesion does not explain why it diagnosed a case as acne and not rosacea. Moreover, such complex tasks require that thorough explanation validation be done by domain experts, which is a time consuming and expensive process. Current dermatological datasets focus either solely on disease diagnosis, or on lesion segmentation (DermNetNZ, 2021; Sun et al., 2016; Tschandl et al., 2018). Having access to expert-annotated dermatological diagnosis explanations would improve the validation of explainability methods and allow the training of intrinsically explainable models. However, to the best of our knowledge, no such dataset exists.

Our contributions are twofold. First, to enable a quantitative assessment of the explainability of dermatological diagnosis models, we introduce DermXDB, a dermatological explainability dataset with gold standard diagnostic explanations provided by eight board-certified dermatologists. DermXDB consists of 554 images from DermNetNZ (2021) and SD-260 (Sun et al., 2016) associated with one of six diagnoses and their explanations in the form of skin lesion characteristics, as defined by Nast et al. (2016). This labeling procedure mimics clinical practice, where dermatologists assess the characteristics of skin lesions to derive and support a tentative diagnosis (Oakley, 2017). An annotation example can be seen in Fig. 2.

Second, we introduce DermX – a novel, clinically-inspired ConvNet architecture for skin disease diagnosis and explanations. This architecture is illustrated in Fig. 1. Following the clinical approach of explaining dermatological diagnoses through skin lesion characteristics, DermX first identifies relevant characteristics in the image (which can also be interpreted as diagnosis explanations), and then relies on them, alongside the image features, to diagnose the case. Using Grad-CAM (Selvaraju et al., 2017), we then localize the predicted characteristics in the image. We validate the plausibility and faithfulness of our explanations using DermXDB as the gold standard for explanations.[1]

1.1. Related work

Machine learning-based dermatological diagnosis systems have been widely investigated, achieving results on par with human experts (Esteva et al., 2017; Tschandl et al., 2020; Jain et al., 2021). These advances in the automated diagnosis of skin lesions were made possible in part by the emergence of various dermatological datasets, which contain images diagnosed by medical experts (Tschandl et al., 2018; DermNetNZ, 2021; Sun et al., 2016). The widely used ISIC dataset (Tschandl et al., 2018) also includes lesion segmentations that can partially serve as a basis for objective explanation measurement. However, these segmentations were not collected to explain the diagnosis, but rather to localize the lesions. This shortcoming becomes critical in diseases such as actinic keratosis, where the area surrounding the lesion is just as important for the diagnosis as the lesion itself (Tschandl et al., 2020).

Explainability is an important topic in machine learning in general and in medical imaging in particular. Saliency-based explainability methods, e.g. Grad-CAM (Selvaraju et al., 2017), are often used as a way to investigate if the models learn relevant features (Tschandl et al., 2020; Zhang et al., 2019; Barata et al., 2021). Other explainability methods, such as LIME (Ribeiro et al., 2016), Kernel-SHAP (Lundberg and Lee, 2017), and Sharp-LIME (Graziani et al., 2021) are less commonly used in the medical imaging literature.

Two works, one in natural language processing and the other in dermatological imaging, have a similar approach to explainability as ours. Within natural language processing, Mathew et al. (2021) propose a framework that explains hateful speech identification. Human readers were asked to identify the most important tokens in a sentence for the prediction of hateful speech. Then, the explanation plausibility and faithfulness of the model-generated explanations were quantified by comparing to the human annotations. Within dermatological image analysis, Barata et al. (2021) investigate how hierarchical taxonomies for skin lesion classification can be used to improve ConvNet skin cancer diagnosis capabilities. They train networks to follow the hierarchical classification of diseases in their prediction, and to focus on relevant parts of the image.

[1] The DermXDB dataset and the implementation of DermX and DermX+ are available at https://github.com/ralucaj/dermx.

Table 1
Distribution of images over DermNetNZ and SD-260, and over the six possible diagnoses.

	Acne	Actinic keratosis	Psoriasis	Seborrheic dermatitis	Viral warts	Vitiligo	Total
DermNetNZ	58	48	47	15	46	77	291
SD-260	61	43	51	79	20	9	263
Total	119	91	98	94	66	86	554

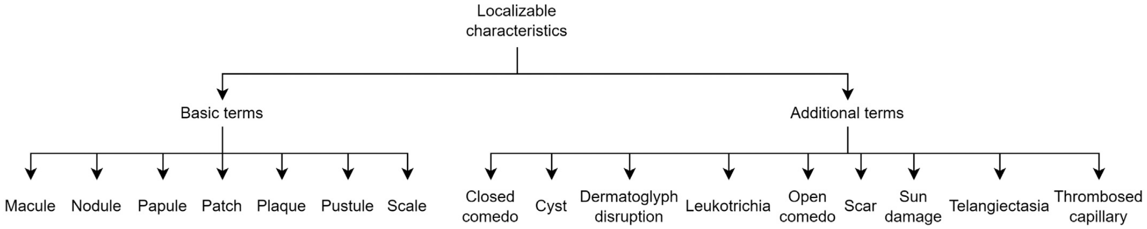


Fig. 3. Localizable characteristics taxonomy. All characteristics were tailored to the six DermXDB diseases using medical resources (Nast et al., 2016; Oakley, 2017), and with the help of two senior dermatologists.

In this work, we combine the two approaches by detecting diagnosis-explaining characteristics, each with its own localization, and train two ConvNets to focus on the relevant part of the image for each characteristic. Both networks are evaluated for the plausibility and faithfulness of their explanations.

2. Material and methods

2.1. Explainability dataset

To enable explainable modeling, we identified the clinically relevant explanation taxonomy, designed an appropriate annotation protocol, and collected expert-labeled data. This resulted in DermXDB: a novel dermatological explainability dataset designed to enable the training of the proposed end-to-end explainable models and quantitative explainability evaluation. The dataset consists of 554 images that belong to one of the following classes: acne, actinic keratosis, psoriasis, seborrheic dermatitis, viral warts, or vitiligo. Images were sourced from DermNetNZ (2021) and SD-260 (Sun et al., 2016) with written permission from the owners. The distribution over datasets and diseases is described in Table 1. All images were evaluated by eight board-certified dermatologists, with between four and twelve years of clinical experience. Each evaluation consists of a diagnosis and supporting explanations in the form of global tags, localizable characteristics, their segmentations, and additional descriptive terms for basic characteristics.

The development of this dataset included several steps. First, we performed multiple experiments to define the target diseases and the nature of the explanations. Second, we selected the six diagnoses and defined the explanation taxonomy illustrated in Fig. 3. Third, the labelers were allowed a short period of time to get accustomed to the annotation protocol and the labeling tool by evaluating images from an internal dataset. Finally, DermXDB images were selected and sent to the dermatologists for labeling.

Preliminary investigation. Nine diseases were initially investigated: psoriasis, rosacea, vitiligo, seborrheic dermatitis, pityriasis rosea, viral warts, actinic keratosis, acne, and impetigo. These diseases were chosen based on prevalence (Lim et al., 2017) and by the expectation that they could be diagnosed using images as the only source of patient information (Oakley, 2017). Dermatologists were asked to diagnose and explain their decision in free-text for over 100 images. During this step, dermatologists could see the original diagnosis of the image, but had the option to disagree with it. This step led to the exclusion of rosacea, impetigo, and pityriasis rosea from future experiments due to the difficulty in diagnosing them in the absence of the patient medical history. It also led to the introduction of a structured ontology for the diagnosis explanations to avoid manual processing of typos and synonyms.

Diagnosis and explanation ontology. Preliminary investigations also highlighted the importance of having a consistent explanation ontology. After analyzing free-text explanations, they were formalized as an extended list of skin lesion characteristics (Nast et al., 2016). The characteristics set was selected to sufficiently explain the six target diseases (Oakley, 2017). With the help of two senior dermatologists, several other relevant characteristics were added.

The resulting set of characteristics was split into non-localizable characteristics (e.g. age or sex), localizable characteristics (e.g. plaque or open comedo), and additional descriptive terms (e.g. red or well-circumscribed), according to the International League of Dermatological Societies' classification (Nast et al., 2016). Fig. 3 illustrates the final DermXDB explanation taxonomy, while more information about the other two types of labels is available in Appendix Figs. A.11 and A.12.

Annotation protocol. Dermatologists were first asked to diagnose the image, and then tag it with characteristics that explain their diagnosis. No information about the gold standard diagnosis or the disease distribution was made available. If the dermatologists were unable to evaluate the image due to poor quality, or if the image depicted a different disease than the target conditions, they had the option to discard it.

Dermatologists could then select diagnosis-supporting non-localizable characteristics as global image tags. Afterwards, they could select and outline localizable characteristics. Dermatologists were instructed to highlight all relevant areas for each characteristic, and were only allowed to include irrelevant areas if separating them from the characteristic was too time consuming or difficult. In other words, they were instructed to favor sensitivity over specificity. Finally, basic terms (as defined in Fig. 3) could be enriched with additional descriptive terms when required for the diagnosis explanation. Once all tags and characteristics were added, the image could be marked as complete.

After the taxonomy and annotation protocol were defined, all dermatologists underwent two rounds of on-boarding in Darwin, a browser-based labeling tool (V7-Labs, 2021). A screenshot of the labeling interface is shown in Appendix Fig. A.13. Following this, they were asked to annotate a set of 630 images from the DermNetNZ and SD-260 datasets.

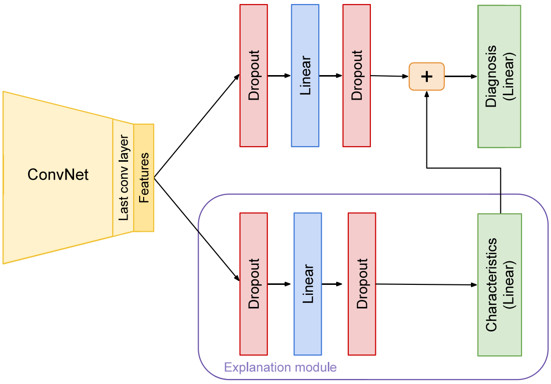


Fig. 4. DermX architecture for image diagnosis with explanations in the form of skin lesion characteristics. The model is trained to predict both diagnoses and characteristics. Image features go through a dimensionality reduction linear layer to ensure that the characteristics are not overshadowed by the image features. The explainability module identifies diagnosis explanations in the form of characteristics, and their localization on the image can be detected through Grad-CAMs.

Data cleaning. Once annotations were performed, the dataset went through two cleanup steps. First, to avoid ambiguities in the dataset, annotations with diagnoses outside the target conditions were discarded. This resulted in 33 images being removed from the dataset because all eight dermatologists tagged these as 'other disease', e.g. acne keloidalis nuchae. The second step was to manually group images from the same patients. For all patients with more than one image, only the first image based on alphabetical order was kept. After cleanup, 554 images were left. Out of all evaluations performed on these images, 150 were discarded due to reports of low image quality, resulting in 4202 individual evaluations.

2.2. Explainable models

We propose two inherently explainable models for joint prediction of diagnosis and explanations. First, we design DermX: an end-to-end clinically-inspired architecture for explainable diagnosis, and train it on the reference diagnosis and expert-identified explanation labels. Next, we build an enhanced model that also includes learning of the explanation localization – DermX+. In the following, we provide a detailed description of each of the models.

DermX model. We propose a clinically-inspired model trained using the data described above. Following the multi-task learning paradigm, the model learns how to predict a diagnosis and its supporting characteristics at the same time. Using a ConvNet as an image feature extractor, we flatten and pass these features into the two prediction modules. The explainability module passes the features through a dense block, composed of a dropout layer, a linear layer with ReLU activations, and another dropout layer. This output is then passed into a linear layer with ten neurons and a logistic function is applied to each to give the probabilistic multi-label predictions, i.e. multiple characteristics can be predicted at the same time. The diagnosis module processes the image features using a similar dense block, after which they are concatenated to the characteristic logits. For this module, the dense block also doubles as a dimensionality reduction component, allowing the image features and the characteristics to have the same order of magnitude. The concatenated features are then passed through a linear layer with six neurons, followed by a softmax function to give our single-label prediction head for diagnoses. Fig. 4 illustrates the DermX architecture.

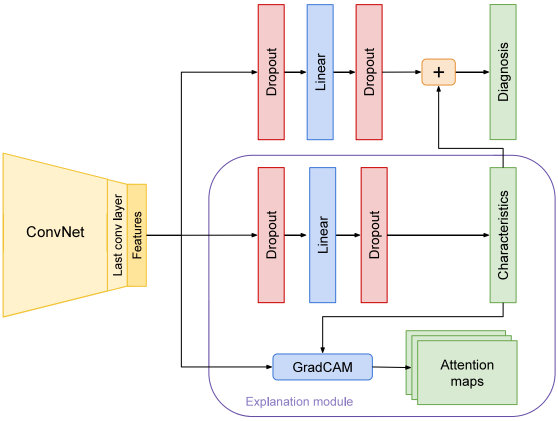


Fig. 5. DermX+ architecture used to generate explanations using guided attention. In addition to the DermX architecture described in Fig. 4, we introduce an additional loss term for the characteristics attention map. The Grad-CAM attention is computed for each predicted characteristic using the features extracted by the last convolutional layer in the backbone network. Characteristic Grad-CAMs are then compared to the downsized fuzzy fusion maps for each characteristic.

DermX optimizes the loss defined as follows. Let $y_{i,d} \in \{0,1\}$ and $z_{i,c} \in \{0,1\}$ be the target diagnosis and target characteristics for image $i \in \{1,\dots,N\}$ in a batch of size N, where $d \in \{1,\dots,D\}$ and $c \in \{1,\dots,C\}$ denote the diagnosis and characteristic class, and let $\hat{y}_{i,d} \in (0,1)$ and $\hat{z}_{i,c} \in (0,1)$ be the diagnosis and characteristics predictions, respectively. The loss can then be written as

$$L = \lambda_D L_D + \lambda_C L_C, \tag{1}$$

where L_D is the categorical cross-entropy diagnosis loss defined as

$$L_D = -\frac{1}{ND}\sum_{i=1}^{N}\sum_{d=1}^{D} y_{i,d}\log \hat{y}_{i,d}, \tag{2}$$

L_C is the binary cross-entropy characteristics loss defined as

$$L_C = -\frac{1}{NC}\sum_{i=1}^{N}\sum_{c=1}^{C}\left(z_{i,c}\log(\hat{z}_{i,c}) + (1-z_{i,c})\log(1-\hat{z}_{i,c})\right), \tag{3}$$

and λ_D and λ_C are hyper-parameters for weighing the relative loss contributions.

DermX+ model. We build on top of the DermX architecture by introducing a guided attention element (Li et al., 2018). Fig. 5 highlights the difference between DermX and DermX+, namely the addition of a characteristic attention component.

In addition to the two losses optimized by DermX and described in Eq. (1), the DermX+ model also optimizes the attention loss term L_A:

$$L = \lambda_D L_D + \lambda_C L_C + \lambda_A L_A, \tag{4}$$

where L_A is the Dice loss for attention

$$L_A = \frac{1}{NC}\sum_{i=1}^{N}\sum_{c=1}^{C}\left(1 - \frac{2A_{i,c}M_{i,c}}{A_{i,c}+M_{i,c}}\right), \tag{5}$$

with $A_{i,c}$ being the attention map, and $M_{i,c}$ being the fuzzy localization label, both for image i and characteristic c.

2.3. Model training and validation

Data. Given the limited size of the dataset, we create a stratified ten-fold cross-validation setup to train explainable models, leading

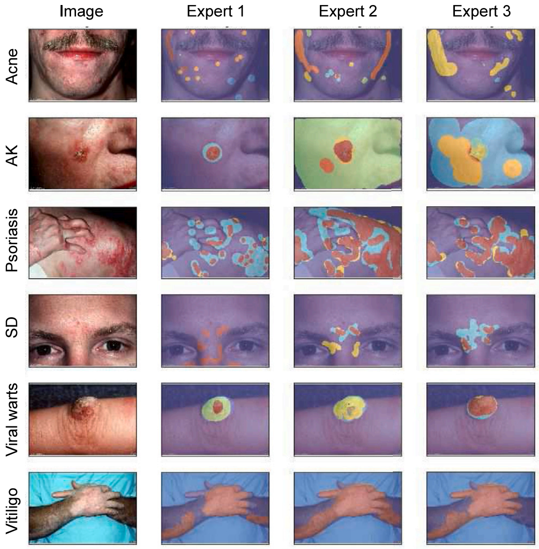


Fig. 6. Differences between dermatologist-labeled attention maps, distributed over the six diseases: acne, actinic keratosis (AK), psoriasis, seborrheic dermatitis (SD), viral warts, and vitiligo. The maps were computed as the union of all characteristics labeled by each of the first three dermatologists. Each color represents a different supporting characteristic.

to approximately 500 training images and 50 test images for each fold. Results presented in this paper are aggregated over all ten folds. For diagnosis prediction we use the gold standard diagnosis label, as defined in the source datasets. A characteristic was marked as relevant for a diagnosis if at least one dermatologist included the characteristic in their decision explanation. Characteristic labels for localization were created as aggregated fuzzy maps, i.e. each pixel value in a mask was generated as a fraction of how many dermatologists included it in their characteristic localization. Only characteristics selected for the correct diagnosis with regard to the gold standard were included both in defining the presence of a characteristic and in the fuzzy map aggregation. This way, we avoid introducing noise due to a mismatch between the diagnosis a dermatologist was explaining and the diagnosis label used to train the network. Additionally, we exclude characteristics that appear in fewer than 30 samples throughout the dataset and characteristics with an inter-rater F1 score below 0.30. We thus focus on closed comedo, dermatoglyph disruption, open comedo, papule, patch, plaque, pustule, scale, scar, and sun damage.

Implementation details. In all experiments, we use an EfficientNet-B2 (Tan and Le, 2019) ConvNet pre-trained on the ImageNet image recognition dataset (Deng et al., 2009) for feature extraction, with all layers fine-tuned on the DermXDB data. Both models were trained for 93 epochs using the AdamW optimizer (Loshchilov and Hutter, 2018), the cosine annealing with warm restarts learning rate scheduler (Loshchilov and Hutter, 2016), and a starting learning rate of 0.0005. Within the dense block we use linear layers with 64 neurons, dropout layers with 0.2 probability, and ReLU activations. DermX is trained with $\lambda_D = 1$, $\lambda_C = 1$, while DermX+ uses $\lambda_D = 1$, $\lambda_C = 1$, and $\lambda_A = 10$. Further information about the hyper-parameters used for training and other implementation details can be found in Appendix Table B.12.

2.4. Explainability evaluation

We measure the performance with regard to the image diagnosis of both our dermatologists and our trained models using the F1 score, sensitivity, and specificity. The same metrics are used to quantify the inter-rater agreement on image diagnosis and characteristics selection between dermatologists. The model performance on characteristics is measured with regard to the fuzzy fusion label for characteristics using the same three metrics. F1 score (also known as the Dice-Sørensen coefficient for pixel-level segmentation), sensitivity, and specificity are also used to measure the inter-rater agreement for the localizable

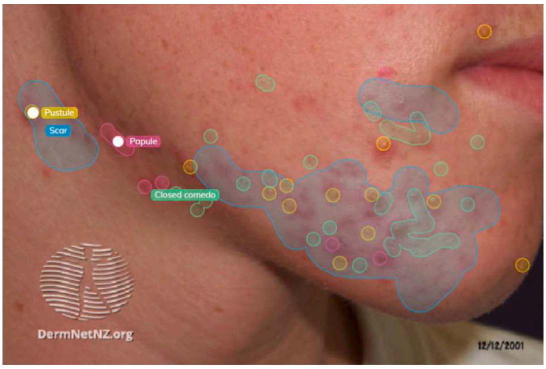


Fig. 7. Characteristics labeled by a dermatologist for an acne case. Following the instructions, no characteristic was segmented but rather the region where they were present was identified without necessarily following the lesion boundaries. For more difficult characteristics to locate, e.g. scars, dermatologists were instructed to brush over entire areas containing the characteristic.

Table 2
Dermatologist inter-rater agreement for the presence or absence of characteristics (mean±std). This analysis shows significant variation in the selection and agreement rates. Characteristics commonly considered important for diagnosing one of the diseases (e.g. comedones, plaques) have higher agreement rates, while uncommon characteristics (e.g. dermatoglyph disruption) display low selection and agreement rates.

	F1 score	Average selection
Dermatoglyph disruption	0.36 ± 0.35	21.50 ± 14.34
Closed comedo	0.54 ± 0.14	41.5 ± 22.20
Open comedo	0.65 ± 0.10	50.88 ± 23.28
Papule	0.67 ± 0.07	138.25 ± 33.44
Patch	0.76 ± 0.11	114.00 ± 43.10
Plaque	0.78 ± 0.06	205.12 ± 35.98
Pustule	0.76 ± 0.06	58.62 ± 13.52
Scale	0.89 ± 0.03	188.50 ± 31.16
Scar	0.46 ± 0.14	41.75 ± 26.84
Sun damage	0.46 ± 0.26	31.62 ± 14.91
Mean	0.63 ± 0.16	86.42 ± 67.05

Table 3
Dermatologist inter-rater localization agreement for localizable characteristics (mean±std). Overlap measures show a significant variation between raters in outlining characteristics. Sensitivity values are high for characteristics that occupy larger areas and that often display well-circumscribed borders (e.g. plaque, scale), but tend to be lower in smaller characteristics (e.g. comedones, pustules).

	F1 score	Sensitivity	Specificity
Dermatoglyph disruption	0.68 ± 0.17	0.82 ± 0.16	0.98 ± 0.03
Closed comedo	0.20 ± 0.21	0.46 ± 0.36	0.89 ± 0.17
Open comedo	0.20 ± 0.18	0.44 ± 0.33	0.91 ± 0.15
Papule	0.27 ± 0.25	0.49 ± 0.34	0.94 ± 0.12
Patch	0.65 ± 0.18	0.80 ± 0.19	0.93 ± 0.11
Plaque	0.64 ± 0.21	0.79 ± 0.21	0.93 ± 0.10
Pustule	0.26 ± 0.17	0.50 ± 0.30	0.98 ± 0.08
Scale	0.51 ± 0.23	0.70 ± 0.25	0.93 ± 0.10
Scar	0.30 ± 0.23	0.56 ± 0.34	0.89 ± 0.14
Sun damage	0.71 ± 0.21	0.84 ± 0.19	0.76 ± 0.22
Mean	0.42 ± 0.20	0.62 ± 0.16	0.91 ± 0.06

Table 4
Comparison of model diagnosis performance with regard to the gold standard, presented as the mean F1 score ± std. The models compared are the diagnosis-only model (Dx), the clinically-inspired diagnosis and characteristics model (DermX), and the DermX model trained with guided attention (DermX+). Dermatologist scores are summarized as mean ± std across the experts. The gold standard is the original image diagnosis as defined by the source dataset.

	Dx	DermX	DermX+	Expert
Acne	0.87 ± 0.05	0.87 ± 0.05	0.86 ± 0.06	0.94 ± 0.02
Actinic keratosis	0.80 ± 0.06	0.79 ± 0.14	0.73 ± 0.10	0.79 ± 0.12
Psoriasis	0.77 ± 0.07	0.73 ± 0.11	0.80 ± 0.09	0.87 ± 0.04
Seborrheic dermatitis	0.77 ± 0.07	0.74 ± 0.09	0.74 ± 0.10	0.75 ± 0.08
Viral warts	0.76 ± 0.18	0.76 ± 0.11	0.76 ± 0.15	0.92 ± 0.05
Vitiligo	0.78 ± 0.10	0.83 ± 0.10	0.86 ± 0.08	0.95 ± 0.02
Mean	0.79 ± 0.05	0.79 ± 0.04	0.79 ± 0.04	0.87 ± 0.08

Table 5
Performance comparison for characteristics identification with regard to dermatologist-generated labels, reported as mean F1 scores. We compare the clinical diagnosis and characteristics model (DermX), the DermX model trained with guided attention (DermX+), and the inter-rater agreement among dermatologists. A characteristic was tagged as present if at least one dermatologist marked it in an image. The F1 score for dermatologists is based on the pairwise inter-rater agreement on characteristics (mean ± std).

	DermX	DermX+	Expert	Samples
Closed comedo	0.76 ± 0.08	0.81 ± 0.12	0.55 ± 0.15	96
Dermatoglyph disruption	0.74 ± 0.20	0.70 ± 0.20	0.37 ± 0.35	54
Open comedo	0.80 ± 0.06	0.80 ± 0.08	0.66 ± 0.10	110
Papule	0.79 ± 0.07	0.80 ± 0.07	0.68 ± 0.07	278
Patch	0.76 ± 0.06	0.79 ± 0.06	0.78 ± 0.11	249
Plaque	0.88 ± 0.03	0.90 ± 0.03	0.79 ± 0.06	352
Pustule	0.79 ± 0.10	0.81 ± 0.06	0.77 ± 0.06	106
Scale	0.79 ± 0.04	0.82 ± 0.05	0.91 ± 0.02	275
Scar	0.78 ± 0.10	0.80 ± 0.08	0.47 ± 0.14	115
Sun damage	0.66 ± 0.11	0.64 ± 0.15	0.46 ± 0.27	78
Mean	0.77 ± 0.03	0.79 ± 0.03	0.64 ± 0.16	171.30

Table 6
DermX performance for characteristics localization with regard to the fuzzy dermatologist localization maps, reported as mean soft sensitivity, specificity, and F1 score. DermX performance metrics are computed only on samples where both the model and the dermatologists agree on the relevance of a characteristic, in order to decouple localization performance from the identification performance.

	Sensitivity	Specificity	F1 score	Samples
Closed comedo	0.69 ± 0.09	0.69 ± 0.05	0.40 ± 0.04	75
Dermatoglyph disruption	0.69 ± 0.09	0.69 ± 0.05	0.28 ± 0.06	36
Open comedo	0.69 ± 0.06	0.68 ± 0.04	0.36 ± 0.05	90
Papule	0.63 ± 0.08	0.72 ± 0.05	0.34 ± 0.04	219
Patch	0.57 ± 0.07	0.78 ± 0.04	0.43 ± 0.05	188
Plaque	0.65 ± 0.06	0.75 ± 0.04	0.43 ± 0.03	314
Pustule	0.69 ± 0.07	0.69 ± 0.05	0.24 ± 0.05	88
Scale	0.65 ± 0.07	0.76 ± 0.04	0.41 ± 0.03	222
Scar	0.64 ± 0.06	0.72 ± 0.05	0.46 ± 0.05	90
Sun damage	0.44 ± 0.08	0.87 ± 0.04	0.56 ± 0.06	50
Mean	0.64 ± 0.02	0.74 ± 0.01	0.39 ± 0.02	137.20

characteristics region outlining overlap. All values are reported as the mean and the standard deviation (std) over the 10 folds.

We define the explainability of our models as having two components: plausibility and faithfulness. For plausibility, we focus on both the identification and the localization of characteristics. First, we measure the F1 score, sensitivity, and specificity per characteristic to measure the models' ability to correctly identify the right explanations. Similar to Mathew et al. (2021), we compare the Grad-CAM activations per characteristic with the fuzzy attention maps for each characteristic, and measure their similarity using the F1 score, sensitivity, and specificity. All pixel-based metrics are implemented using fuzzy logic, as follows:

$$\mathrm{F1} = \frac{2\sum_{p\in P}\min(A_p, M_p)}{\sum_{p\in P}(A_p) + \sum_{p\in P}(M_p)}, \tag{6}$$

Table 7
DermX+ performance for characteristics localization with regard to the fuzzy dermatologist localization maps, reported as mean soft sensitivity, specificity, and F1 score. DermX+ values are computed only on samples where both the model and the dermatologists agree on the relevance of a characteristic, in order to decouple localization performance from the identification performance.

	Sensitivity	Specificity	F1 score	Samples
Closed comedo	0.11 ± 0.08	0.96 ± 0.02	0.10 ± 0.09	63
Dermatoglyph disruption	0.60 ± 0.15	0.97 ± 0.02	0.60 ± 0.12	34
Open comedo	0.06 ± 0.05	0.97 ± 0.02	0.06 ± 0.05	93
Papule	0.19 ± 0.06	0.97 ± 0.02	0.18 ± 0.04	232
Patch	0.56 ± 0.06	0.89 ± 0.03	0.53 ± 0.05	191
Plaque	0.65 ± 0.06	0.91 ± 0.01	0.61 ± 0.05	312
Pustule	0.04 ± 0.05	0.98 ± 0.01	0.03 ± 0.03	91
Scale	0.58 ± 0.11	0.89 ± 0.03	0.49 ± 0.07	224
Scar	0.35 ± 0.15	0.92 ± 0.04	0.35 ± 0.12	93
Sun damage	0.53 ± 0.19	0.90 ± 0.10	0.58 ± 0.20	47
Mean	0.37 ± 0.01	0.94 ± 0.00	0.36 ± 0.01	138.00

$$\mathrm{Sensitivity} = \frac{\sum_{p\in P}\min(A_p, M_p)}{\sum_{p\in P}(M_p)}, \tag{7}$$

$$\mathrm{Specificity} = \frac{\sum_{p\in P}\min(1 - A_p, 1 - M_p)}{\sum_{p\in P}(1 - M_p)}, \tag{8}$$

where P represents the pixels included in the analysis, A defines the class activations, and M represents the fuzzy label maps.

Following the comprehensiveness evaluation described by DeYoung et al. (2020), we measure the faithfulness of our models through the use of contrastive examples. Given a model m, an input image x, a set of explanation outlines e, a contrastive image x_e where all areas marked as an explanation for the image x were occluded, and the class probability output $m(x)$ for the predicted class on the original input x we measure the faithfulness F as

$$F = m(x) - m(x_e). \tag{9}$$

In other words, the faithfulness describes what impact removing the explanations e from the image would have on the decision of model m. We decided not to include the sufficiency metric as it would lead to out-of-distribution images, such as a blank background with a plaque or a couple of pustules.

Finally, given the intrinsic disagreement between experts within medical fields, we postulate that explainable models should be able to properly argue their decisions, regardless of whether it matches the gold standard or not. Similar to how dermatologists may debate the correct diagnosis for a case by highlighting different explanations that support their decision, we expect an explainable model to do the same. However, as we do not always have the gold standard explanation for a wrong diagnosis, we need to define a basic set of explanations for any disease. To this end, we define the expected explanation as the prevalence of each characteristic within the dermatologists explanations for a diagnosis (Appendix Table A.11). Then, for the wrongly predicted diagnoses we compare the set of characteristics associated with that prediction with the expected explanation for the predicted diagnosis. For example, a case incorrectly classified as psoriasis is expected to be explained using one or several of papule, plaque, and scale, which are commonly used by dermatologists in their explanations of psoriasis. We evaluate how the model explanation for wrong diagnoses by computing the precision of the model's explanations with regard to the expected explanation for a diagnosis.

3. Results

3.1. DermXDB analysis

We first analyzed the data focusing on dermatologist performance with regard to the gold standard diagnosis and their inter-rater agreement on both diagnoses and supporting characteristics. A total of 554

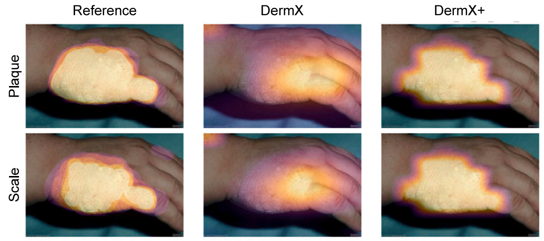


Fig. 8. Characteristic attention maps for a correctly classified psoriasis case, for the two identified characteristics: plaque (first row) and scale (second row). The first column shows the dermatologist-derived fuzzy attention map, the second one illustrates the Grad-CAM for each characteristic generated by DermX, while the last column shows the DermX+ Grad-CAM maps. DermX+ displays much closer results to the gold standard, while DermX maps include more irrelevant information, such as the finger knuckles.

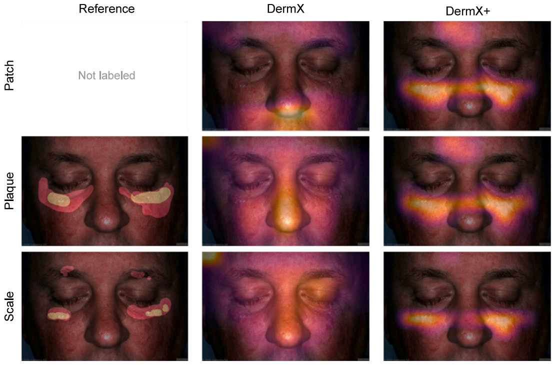


Fig. 9. Characteristic attention maps for a psoriasis case wrongly classified as seborrheic dermatitis by both DermX and DermX+. While both models detect plaque and scale in the image, their detection of patch, a characteristic indicative of seborrheic dermatitis, leads them to misdiagnose the image.

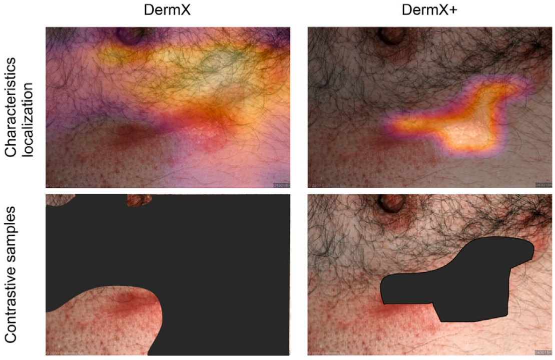


Fig. 10. Contrastive samples on a DermXDB psoriasis image for DermX and DermX+. DermX, a more sensitive and less specific model, occludes large parts of the image, while DermX+, a more specific and less sensitive model, occludes only the lesion. When evaluating the contrastive sample for DermX+, the model has the possibility to use other diagnosis hints in the image (defined in DermXDB as non-localizable characteristics) that are occluded in the DermX sample.

images were included in this analysis, each with eight evaluations performed by board-certified dermatologists. The dermatologist diagnostic performance in terms of mean F1 score with regard to the gold standard varies between 0.75 for seborrheic dermatitis and 0.95 for vitiligo. Aggregated F1 scores can be seen in Table 4. A full description of the dermatologist performance with regard to the gold standard is available in Appendix Table A.8.

Inter-rater agreement on characteristics, as described in Table 2, varies significantly more, partially due to the lower number of selections per class. Most basic terms display high levels of agreement, with F1 scores between 0.67 and 0.89. The exceptions are macule with an F1 score of 0.12 and nodule with an F1 score of 0.17, both also displaying low selection rates. Several additional terms, such as open and closed comedones, display levels of agreement similar to the basic terms. Fig. 6 illustrates an example of disagreement between three dermatologists on the location of supporting characteristics on one random case for each disease, while Fig. 7 highlights how the protocol was followed by a dermatologist in an acne case. Additional metrics for the full set of characteristics are described in Appendix Table A.9.

Outlining characteristics is a more difficult task, as confirmed by the low inter-rater F1 scores reported in Table 3. The lower F1 values can also be explained by how difficult outlining small or poorly circumscribed characteristics is. In terms of sensitivity, we notice the same trend as in binary agreement: dermatologists tend to agree more on the basic terms. Metrics for the full set of localizable characteristics are presented in Appendix Table A.10.

3.2. Explainable model

We trained a clinically-inspired model from Fig. 1 (DermX), and the same model architecture trained with guided attention (DermX+) for characteristics localization. We also train a diagnosis-only model (Dx) to check whether adding explanations impacts the diagnosis performance of DermX and DermX+.

Table 4 compares the diagnostic performance between all three models and the dermatologists with regard to the gold standard diagnosis. More information about their diagnostic performance is presented in Appendix B. For comparison, we trained a diagnosis-only model with a ResNet50 (He et al., 2016) base to validate the choice of architecture, and a diagnosis-only model trained with proportional class weights. The ResNet-based model achieved a macro F1-score of 0.79 ± 0.06, while the weighted class model showed a similar macro F1-score of 0.78±0.05. More information about these two models is available in Appendix Table B.13. Additionally, we trained four interpretable models on the characteristics data for diagnosis prediction: a logistic regression model, a decision tree, a k-nearest neighbor with five neighbors, and a categorical naive Bayes models. These models obtained macro F1 scores of 0.86 ± 0.04, 0.85 ± 0.05, 0.80 ± 0.05, and 0.86 ± 0.05, respectively.

All models display similar F1 scores on all six diseases. The best results are obtained for vitiligo and acne, two disease classes where dermatologists also display high F1 score values. Seborrheic dermatitis on the other hand seems to be a difficult disease class for both dermatologists and models. For the rest of the results section we will focus on DermX and DermX+.

In terms of explanation plausibility, we look at both the identification of explanations, defined as the ability to detect the same characteristics as a dermatologist, and at their localization in the image. A comparison of F1 scores is described in Table 5. The two models perform well for explanation identification, with DermX+ obtaining slightly better results on most characteristics. Compared to dermatologists, the models perform within standard deviation bounds of the inter-rater agreement. Additional metrics are reported in Appendix Tables B.16 and B.17.

The localization plausibility of the models' explanation is quantified in Tables 6 and 7, with more statistics being presented in Appendix Tables B.18, B.19, B.20, and B.21. DermX performs adequately well on all characteristics. DermX+ is better at localizing large characteristics, e.g. patches or scales, but performs poorly on smaller characteristics, e.g. open and closed comedones. Dermatologists F1 scores indicate

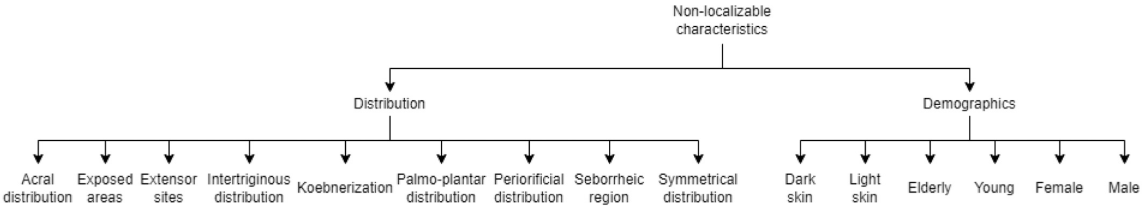


Fig. A.11. Non-localizable characteristics taxonomy. These characteristics were added to the International League of Dermatological Societies' classification as global image tags after being flagged as relevant by our senior dermatologists.

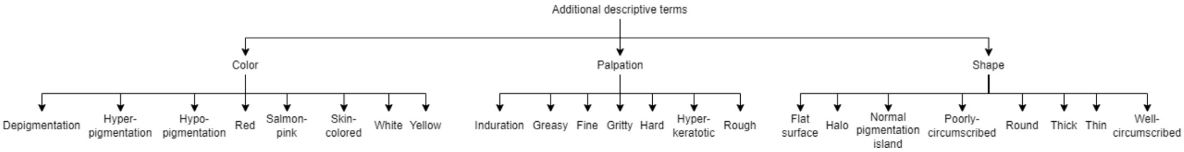


Fig. A.12. Additional descriptive terms for localizable characteristics. All terms were tailored for the six diseases from medical resources (Nast et al., 2016; Oakley, 2017), and with the help of two senior dermatologists.

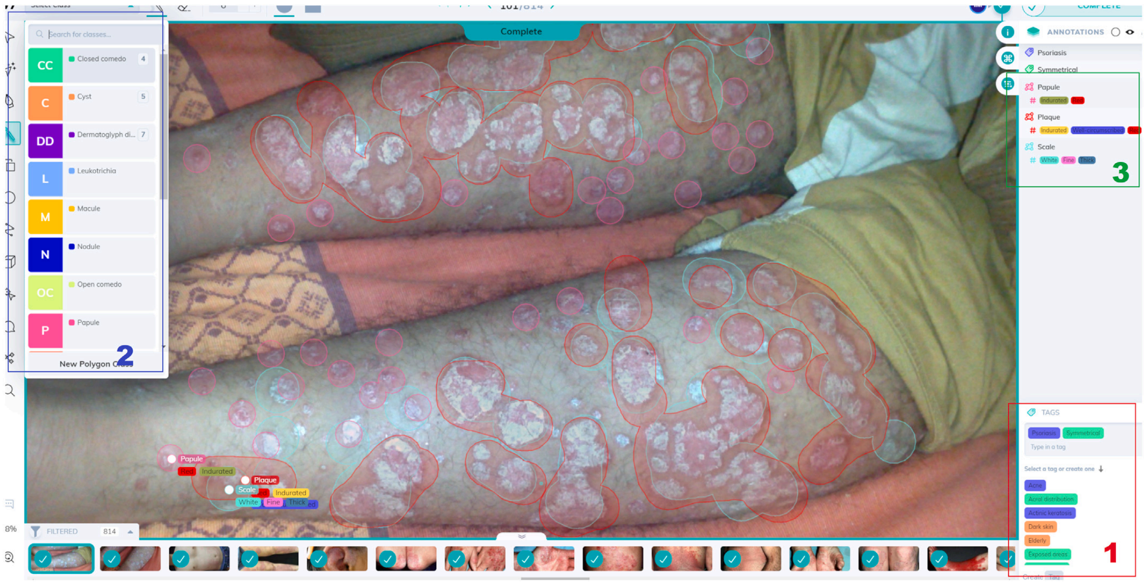


Fig. A.13. Labeling tool interface, exemplified for a psoriasis case from the SD-260 dataset. In the global tag search box (area 1, bottom right), dermatologists can select the disease, relevant demographics information, and lesion distribution. The brush selection menu (area 2, top left) allows them to select and mark localizable characteristics on the image. The full annotation menu (area 3, top right) is used to select of additional descriptive terms for the localized basic terms.

Table A.8
Dermatologist diagnosis performance with regard to the gold standard (mean±std).

	F1 score	Sensitivity	Specificity	Average selection
Acne	0.94 ± 0.02	0.90 ± 0.04	0.99 ± 0.00	102.75 ± 10.00
Actinic keratosis	0.79 ± 0.12	0.68 ± 0.16	1.00 ± 0.00	57.62 ± 16.54
Psoriasis	0.87 ± 0.04	0.88 ± 0.04	0.97 ± 0.02	98.00 ± 8.09
Seborrheic dermatitis	0.75 ± 0.08	0.64 ± 0.11	0.99 ± 0.01	60.38 ± 9.23
Viral warts	0.92 ± 0.05	0.85 ± 0.08	1.00 ± 0.00	55.38 ± 5.05
Vitiligo	0.95 ± 0.02	0.90 ± 0.05	1.00 ± 0.00	76.25 ± 4.82
Mean	0.87 ± 0.08	0.81 ± 0.11	0.99 ± 0.01	75.06 ± 19.15

that the two models are, in some characteristics, within standard deviation of the inter-rater agreement. For other characteristics, such as dermatoglyph disruption for DermX and pustule for DermX+, the model performance is below the expert inter-rater agreement. Fig. 8 illustrates the explanations given for a correctly predicted psoriasis case by DermX and DermX+, respectively, while Fig. 9 shows the explanations given by the two ConvNets for a misclassified psoriasis case.

Explanation precision scores for the correct diagnosis prediction were computed with regard to the dermatologist labels. The resulting values are 0.88 ± 0.03 for DermX and 0.90 ± 0.03 for DermX+. On the wrong diagnosis prediction, DermX precision is 0.85 ± 0.06, while DermX+ precision is 0.86±0.04. Mean faithfulness results are 0.42±0.06 for DermX and 0.27 ± 0.06 for DermX+.

Table A.9
Dermatologist inter-rater agreement for the presence or absence of characteristics (mean±std). This analysis shows significant variation in the selection and agreement rates. Characteristics commonly considered important for diagnosing one of the diseases (e.g. comedones, plaques) have higher agreement rates, while uncommon characteristics (e.g. leukotrichia, telangiectasia) display low selection and agreement rates.

	F1 score	Sensitivity	Specificity	Cohen's kappa	Average selection
Basic terms					
Macule	0.12 ± 0.15	0.18 ± 0.23	0.95 ± 0.07	0.09 ± 0.15	25.25 ± 30.37
Nodule	0.17 ± 0.16	0.26 ± 0.27	0.97 ± 0.04	0.16 ± 0.15	17.88 ± 17.42
Papule	0.67 ± 0.07	0.69 ± 0.13	0.86 ± 0.08	0.52 ± 0.09	138.25 ± 33.44
Patch	0.76 ± 0.11	0.79 ± 0.18	0.92 ± 0.08	0.68 ± 0.15	114.00 ± 43.10
Plaque	0.78 ± 0.06	0.80 ± 0.11	0.84 ± 0.08	0.62 ± 0.09	205.12 ± 35.98
Pustule	0.76 ± 0.06	0.78 ± 0.11	0.96 ± 0.02	0.73 ± 0.06	58.62 ± 13.52
Scale	0.89 ± 0.03	0.90 ± 0.06	0.93 ± 0.04	0.82 ± 0.04	188.50 ± 31.16
Additional terms					
Closed comedo	0.54 ± 0.14	0.64 ± 0.25	0.96 ± 0.04	0.51 ± 0.14	41.5 ± 22.20
Cyst	0.20 ± 0.16	0.31 ± 0.34	0.99 ± 0.01	0.19 ± 0.16	6.25 ± 6.70
Dermatoglyph disruption	0.36 ± 0.35	0.39 ± 0.38	0.97 ± 0.03	0.37 ± 0.34	21.50 ± 14.34
Leukotrichia	0.43 ± 0.41	0.45 ± 0.43	1.00 ± 0.01	0.45 ± 0.41	4.62 ± 2.69
Open comedo	0.65 ± 0.10	0.71 ± 0.21	0.96 ± 0.04	0.61 ± 0.11	50.88 ± 23.28
Scar	0.46 ± 0.14	0.57 ± 0.26	0.95 ± 0.05	0.42 ± 0.14	41.75 ± 26.84
Sun damage	0.46 ± 0.26	0.53 ± 0.29	0.97 ± 0.02	0.43 ± 0.25	31.62 ± 14.91
Telangiectasia	0.17 ± 0.25	0.19 ± 0.28	0.99 ± 0.01	0.19 ± 0.25	6.12 ± 5.60
Thrombosed capillaries	0.36 ± 0.27	0.45 ± 0.37	0.98 ± 0.02	0.35 ± 0.26	15.88 ± 11.78
Mean	0.49 ± 0.24	0.54 ± 0.22	0.95 ± 0.04	0.45 ± 0.21	60.48 ± 62.97

Table A.10
Dermatologist inter-rater localization agreement for localizable characteristics (mean±std). Overlap measures show a significant variation between raters in outlining characteristics. Sensitivity values are high for characteristics that occupy larger areas and that often display well-circumscribed borders (e.g. plaque, scale), but tend to be lower in smaller characteristics (e.g. comedones, pustules).

	F1 score	Sensitivity	Specificity
Basic terms			
Macule	0.21 ± 0.16	0.44 ± 0.3	0.95 ± 0.11
Nodule	0.31 ± 0.24	0.55 ± 0.33	0.96 ± 0.09
Papule	0.27 ± 0.25	0.49 ± 0.34	0.94 ± 0.12
Patch	0.65 ± 0.18	0.80 ± 0.19	0.93 ± 0.11
Plaque	0.64 ± 0.21	0.79 ± 0.21	0.93 ± 0.10
Pustule	0.26 ± 0.17	0.50 ± 0.30	0.98 ± 0.08
Scale	0.51 ± 0.23	0.70 ± 0.25	0.93 ± 0.10
Additional terms			
Closed comedo	0.20 ± 0.21	0.46 ± 0.36	0.89 ± 0.17
Cyst	0.39 ± 0.27	0.59 ± 0.31	0.98 ± 0.07
Dermatoglyph disruption	0.68 ± 0.17	0.82 ± 0.16	0.98 ± 0.03
Leukotrichia	0.50 ± 0.14	0.70 ± 0.21	0.96 ± 0.07
Open comedo	0.20 ± 0.18	0.44 ± 0.33	0.91 ± 0.15
Scar	0.30 ± 0.23	0.56 ± 0.34	0.89 ± 0.14
Sun damage	0.71 ± 0.21	0.84 ± 0.19	0.76 ± 0.22
Telangiectasia	0.29 ± 0.14	0.53 ± 0.28	0.93 ± 0.12
Thrombosed capillaries	0.42 ± 0.22	0.65 ± 0.28	0.99 ± 0.02
Mean	0.41 ± 0.18	0.62 ± 0.14	0.93 ± 0.05

4. Discussion

To the best of our knowledge, DermX is the first end-to-end framework created for the purpose of explaining automated dermatological diagnoses. The two ConvNets we introduce, DermX and DermX+, mimic the dermatological approach to diagnosing skin conditions: first they recognize supporting characteristics, then they use these characteristics as well as other high level information to arrive at a diagnosis. In addition to identifying supporting characteristics as explanations to a diagnosis, DermX+ also learns the localization of the explanations via the guided attention loss. The decision to use an attention mechanism for localization rather than a semantic segmentation approach was guided by the design of the annotation protocol. Because dermatologists were instructed to highlight explanation regions in an image with a focus on sensitivity instead of specificity, the resulting outlines are not well suited as segmentation masks. For this work to be possible, we collected diagnoses and supporting characteristics for 554 images from eight board-certified dermatologists.

During the process of collecting the DermXDB data, we found that dermatologists often focus on different characteristics when diagnosing a case. While most explanations for diseases display a set of common characteristics, such as scales, plaques, and papules for psoriasis, there is also a long tail of relevant characteristics that are not always selected. In addition, we found that inter-rater agreement was low for characteristics localization. This may be caused by the difficulty in outlining characteristics with poorly defined boundaries, such as patch, but also by dermatologists differing in their approach to outlining smaller characteristics, such as open and closed comedones.

The contrast between high agreement on diagnoses and low agreement on supporting characteristics illustrates how different experts perceive explanations in different ways. Although they generally agree on the diagnosis, dermatologists focus on different characteristics to explain their decision. To properly evaluate a model's explanations, we must therefore consider the opinions of multiple experts. Moreover, this intrinsic variability in how experts approach explanations lends more urgency to the need for quantifiable explanation methods.

From a modeling perspective, our results contradict the common adage that there must be a trade-off between predictive power and explainability. DermX and DermX+ both report the same diagnosis performance as a standard diagnosis-only ConvNet, while also offering plausible explanations for their decisions. Even in cases where they predict the wrong diagnosis, both models provide arguments that make sense for their prediction. Most explanations given by both models are within standard deviation of the inter-rater agreement on characteristics, suggesting that either model may function as a second opinion with realistic decision explanations.

When compared to interpretable models trained on the characteristics data, both DermX and DermX+ obtain a diagnosis performance within standard deviation of the models using manually labeled features. None of the models we trained obtains a diagnosis performance as high as that of experts. We postulate that this is due to the difficulty of the dataset, as shown by the inter-rater agreement in Table 4, and due to the limited amount of training data. On the other hand, our results are on par with the diagnosis accuracy reported by other research groups using dermatological clinical photography, which varies between 56.7% on 134 classes (Han et al., 2020) and 86.53% on four classes (Burlina et al., 2019).

Our localization results for both models are lower than the inter-rater agreement on expert-derived maps for most characteristics. This may in part be due to the low inter-rater agreement on the localization data, and in part due to the small scale at which the maps were

Table A.11
Characteristics prevalence per disease.

	Acne	Actinic keratosis	Psoriasis	Seborrheic dermatitis	Viral warts	Vitiligo
Closed comedo	0.40	0.00	0.00	0.00	0.00	0.00
Dermatoglyph disruption	0.00	0.00	0.00	0.00	0.38	0.00
Open comedo	0.49	0.00	0.00	0.01	0.00	0.00
Papule	0.70	0.13	0.18	0.04	0.65	0.00
Patch	0.02	0.32	0.02	0.24	0.00	0.97
Plaque	0.04	0.59	0.96	0.71	0.46	0.01
Pustule	0.56	0.00	0.01	0.00	0.00	0.00
Scale	0.01	0.81	0.89	0.78	0.05	0.00
Scar	0.38	0.03	0.00	0.00	0.00	0.00
Sun damage	0.00	0.47	0.01	0.01	0.00	0.01

Table B.12
Training hyper-parameters common to all trained models.

Hyperparameter	Value
Batch size	32
Rotation	10
Zoom	0.15
Brightness	0.35
Contrast	0.20
Saturation	0.20
Scale	(0.85, 1.15)
Translate	(0.15, 0.15)
Hue	0.15
Dropout	0.20

computed (nine by nine pixels for the EfficientNet-B2 architecture). However, the high sensitivity values show that these maps are often good enough to give a visual hint as to the location of the characteristic in an image. Such a hint would be useful in cases where an expert using DermX or DermX+ as a second opinion did not notice that characteristic. Comparing the two models, DermX+ displays lower overall F1 scores than DermX, while showcasing higher overall specificity and high sensitivity on large characteristics. This may be explained by its training target: dermatologist attention maps were linearly scaled down to the size of the feature maps, which may have reduced the target attention map of small characteristics to an almost empty mask. Another possible explanation is DermX's reliance on the sometimes noisy localization data. In particular, for characteristics smaller than 1 cm (closed comedo, open comedo, papule, and pustule), DermX+ is clearly outperformed by DermX due to the lower specificity and higher sensitivity of DermX. In the future, we plan on investigating different ways of downscaling the masks, and to increase the feature map size to take advantage of the high resolution gold standard attention maps.

Mean faithfulness scores above zero for both models prove that the characteristic localizations are indeed explanations about the diagnosis decision mechanisms of the models. DermX+, a more specific model in terms of characteristic localization, has lower faithfulness scores than DermX, which tends to include adjacent regions in its localizations. Fig. 10 showcases the impact a model's specificity and sensitivity have on the contrastive samples, and therefore on the faithfulness metric. In this example, the contrastive sample created by DermX+ still displays image-level non-localizable characteristics, information which is occluded in the DermX contrastive sample. This further confirms the importance of image-level tags in skin lesion diagnosis, e.g. by noticing that acne is predominantly located on the face or upper trunk, or that actinic keratosis most commonly affects elderly people.

This work opens many new research avenues in the domain of medical image diagnosis explainability. From a dermatological data perspective, we plan on adding more diseases and supporting characteristics to DermXDB. The annotation protocol developed as part of DermXDB can serve as an inspiration not only for explaining other dermatological diseases, but also for different radiology and pathology investigations. In radiology, imaging findings are routinely recorded, including the supporting characteristics for the diagnosis. For example, the malignancy of a lesion seen on a mammogram could be supported by localized characteristics such as calcifications and dense tissue (Sickles et al., 2013). The network architectures we proposed could also be applied to learning the supporting radiological findings as explanations to diagnoses provided appropriately labeled datasets. From a modeling perspective, we will focus on leveraging the full potential of DermXDB by adding image-level explanations to the diagnosis models, and by incorporating the additional descriptive terms into the explanation setup. More work can be done in improving the characteristic localization. We will be focusing in particular on introducing the adversarial loss described in Li et al. (2018) for semi-supervised attention guidance. Another approach we will be to train object detection networks (Tan et al., 2020; Redmon and Farhadi, 2018) to detect the supporting characteristics alongside the diagnosis. Once the localization reaches a higher performance, a true test of the DermX architecture would be to set up a clinical trial where its predictions would be used as a second opinion for health care professionals of various levels of expertise.

5. Conclusions

In this work, we introduce DermX – a novel, clinically-inspired explainable ConvNet architecture for skin lesion diagnosis. We also introduce a variation named DermX+ that adds a guided attention loss such that the localization of lesion characteristics becomes a part of the supervised training. We quantify the explanation quality by comparing it with explanations given by board-certified dermatologists with different levels of clinical experience. To facilitate future work, we release this explainability dataset to the public, and describe the annotation protocol used for its creation.

Declaration of competing interest

The authors declare the following financial interests/personal relationships which may be considered as potential competing interests: Raluca Jalaboi reports financial support was provided by Innovation Fund Denmark. Ole Winther reports financial support was provided by Novo Nordisk Foundation. Ole Winther reports financial support was provided by Danish National Research Foundation. Frederik Faye, Mauricio Orbes-Arteaga, Dan Jørgensen, and Alfiia Galimzianova have no conflicts of interest.

Data availability

Data will be made available on request.

Acknowledgments

Raluca Jalaboi's work was supported in part by the Danish Innovation Fund, Denmark under Grant 0153-00154 A. Ole Winther's work was funded in part by the Novo Nordisk Foundation, Denmark through the Center for Basic Machine Learning Research in Life Science (NNF20OC0062606). Ole Winther acknowledges supporting the Pioneer Centre for AI, DNRF, Denmark grant number P1.

Table B.13
Comparison of model diagnosis performance with regard to the gold standard, presented as the mean F1 score ±std. The models compared are the diagnosis-only model (Dx), diagnosis-only with a ResNet50 base (DxRN), a diagnosis-only model trained with class weights, the clinically-inspired diagnosis and characteristics model (DermX), and the DermX model trained with guided attention (DermX+). Dermatologist scores are summarized as mean ±std across the experts. The gold standard is the original image diagnosis as defined by the source dataset.

	Dx	DxRN	DxW	DermX	DermX+	Expert
Acne	0.87 ± 0.05	0.90 ± 0.06	0.85 ± 0.05	0.87 ± 0.05	0.86 ± 0.06	0.94 ± 0.02
Actinic keratosis	0.80 ± 0.06	0.74 ± 0.16	0.77 ± 0.15	0.79 ± 0.14	0.73 ± 0.10	0.79 ± 0.12
Psoriasis	0.77 ± 0.07	0.76 ± 0.07	0.77 ± 0.06	0.73 ± 0.11	0.80 ± 0.09	0.87 ± 0.04
Seborrheic dermatitis	0.77 ± 0.07	0.72 ± 0.11	0.75 ± 0.14	0.74 ± 0.09	0.74 ± 0.10	0.75 ± 0.08
Viral warts	0.76 ± 0.18	0.74 ± 0.16	0.73 ± 0.20	0.76 ± 0.11	0.76 ± 0.15	0.92 ± 0.05
Vitiligo	0.78 ± 0.10	0.85 ± 0.06	0.82 ± 0.07	0.83 ± 0.10	0.86 ± 0.08	0.95 ± 0.02
Mean	0.79 ± 0.05	0.79 ± 0.06	0.78 ± 0.05	0.79 ± 0.04	0.79 ± 0.04	0.87 ± 0.08

Table B.14
DermX diagnostic performance with regard to the gold standard.

	F1 score	Sensitivity	Specificity
Acne	0.87 ± 0.05	0.89 ± 0.07	0.96 ± 0.01
Actinic keratosis	0.79 ± 0.14	0.74 ± 0.14	0.97 ± 0.03
Psoriasis	0.73 ± 0.11	0.77 ± 0.11	0.92 ± 0.05
Seborrheic dermatitis	0.74 ± 0.09	0.76 ± 0.10	0.93 ± 0.04
Viral warts	0.76 ± 0.11	0.68 ± 0.15	0.99 ± 0.01
Vitiligo	0.83 ± 0.10	0.82 ± 0.13	0.97 ± 0.03
Mean	0.79 ± 0.04	0.78 ± 0.04	0.96 ± 0.01

Table B.15
DermX+ diagnosis performance with regard to the gold standard.

	F1 score	Sensitivity	Specificity
Acne	0.86 ± 0.06	0.92 ± 0.06	0.94 ± 0.04
Actinic keratosis	0.73 ± 0.10	0.69 ± 0.15	0.96 ± 0.02
Psoriasis	0.80 ± 0.09	0.83 ± 0.08	0.95 ± 0.04
Seborrheic dermatitis	0.74 ± 0.10	0.76 ± 0.11	0.94 ± 0.03
Viral warts	0.76 ± 0.15	0.70 ± 0.18	0.98 ± 0.01
Vitiligo	0.86 ± 0.08	0.83 ± 0.11	0.98 ± 0.02
Mean	0.79 ± 0.04	0.79 ± 0.04	0.96 ± 0.01

Table B.16
DermX performance on the presence or absence of characteristics with regard to the dermatologist-generated labels.

	F1 score	Sensitivity	Specificity	Samples
Closed comedo	0.76 ± 0.08	0.79 ± 0.01	0.95 ± 0.03	96
Dermatoglyph disruption	0.74 ± 0.20	0.68 ± 0.21	0.99 ± 0.02	54
Open comedo	0.80 ± 0.06	0.82 ± 0.09	0.95 ± 0.02	110
Papule	0.79 ± 0.07	0.79 ± 0.10	0.80 ± 0.08	278
Patch	0.76 ± 0.06	0.76 ± 0.11	0.82 ± 0.06	249
Plaque	0.88 ± 0.03	0.89 ± 0.03	0.74 ± 0.11	352
Pustule	0.79 ± 0.10	0.83 ± 0.14	0.94 ± 0.02	106
Scale	0.79 ± 0.04	0.81 ± 0.07	0.77 ± 0.07	275
Scar	0.78 ± 0.10	0.80 ± 0.14	0.94 ± 0.04	115
Sun damage	0.66 ± 0.11	0.64 ± 0.15	0.96 ± 0.03	78
Mean	0.77 ± 0.03	0.78 ± 0.04	0.88 ± 0.02	171.30

Table B.17
DermX+ performance on the presence or absence of characteristics with regard to the dermatologist-generated labels.

	F1 score	Sensitivity	Specificity	Samples
Closed comedo	0.81 ± 0.12	0.87 ± 0.13	0.94 ± 0.04	96
Dermatoglyph disruption	0.70 ± 0.2	0.64 ± 0.23	0.99 ± 0.01	54
Open comedo	0.80 ± 0.08	0.85 ± 0.09	0.93 ± 0.05	110
Papule	0.80 ± 0.07	0.83 ± 0.07	0.76 ± 0.11	278
Patch	0.79 ± 0.06	0.77 ± 0.11	0.87 ± 0.08	249
Plaque	0.90 ± 0.03	0.89 ± 0.05	0.85 ± 0.06	352
Pustule	0.81 ± 0.06	0.86 ± 0.08	0.94 ± 0.03	106
Scale	0.82 ± 0.05	0.82 ± 0.07	0.83 ± 0.07	275
Scar	0.80 ± 0.08	0.82 ± 0.14	0.94 ± 0.03	115
Sun damage	0.64 ± 0.15	0.60 ± 0.17	0.96 ± 0.03	78
Mean	0.79 ± 0.03	0.80 ± 0.04	0.90 ± 0.02	171.30

Table B.18
DermX localization performance for localizable characteristics (mean±std) with regard to the fuzzy dermatologist attention maps.

	F1 score	Sensitivity	Specificity	Samples
Closed comedo	0.40 ± 0.04	0.69 ± 0.09	0.69 ± 0.05	75
Dermatoglyph disruption	0.28 ± 0.06	0.69 ± 0.09	0.69 ± 0.05	36
Open comedo	0.36 ± 0.05	0.69 ± 0.06	0.68 ± 0.04	90
Papule	0.34 ± 0.04	0.63 ± 0.08	0.72 ± 0.05	219
Patch	0.43 ± 0.05	0.57 ± 0.07	0.78 ± 0.04	188
Plaque	0.43 ± 0.03	0.65 ± 0.06	0.75 ± 0.04	314
Pustule	0.24 ± 0.05	0.69 ± 0.07	0.69 ± 0.05	88
Scale	0.41 ± 0.03	0.65 ± 0.07	0.76 ± 0.04	222
Scar	0.46 ± 0.05	0.64 ± 0.06	0.72 ± 0.05	90
Sun damage	0.56 ± 0.06	0.44 ± 0.08	0.87 ± 0.04	50
Mean	0.39 ± 0.02	0.64 ± 0.02	0.74 ± 0.01	137.20

Table B.19
DermX characteristics localization performance with regard to fuzzy dermatologist attention maps. The results include the localization performance of characteristics identified by the dermatologists but not by the model.

	F1 score	Sensitivity	Specificity	Samples
Closed comedo	0.31 ± 0.04	0.55 ± 0.11	0.76 ± 0.05	96
Dermatoglyph disruption	0.19 ± 0.06	0.48 ± 0.17	0.79 ± 0.07	54
Open comedo	0.30 ± 0.05	0.57 ± 0.08	0.73 ± 0.05	110
Papule	0.27 ± 0.05	0.50 ± 0.08	0.78 ± 0.05	278
Patch	0.33 ± 0.05	0.44 ± 0.08	0.83 ± 0.04	249
Plaque	0.38 ± 0.02	0.58 ± 0.05	0.77 ± 0.04	352
Pustule	0.20 ± 0.05	0.58 ± 0.14	0.74 ± 0.07	106
Scale	0.33 ± 0.03	0.53 ± 0.08	0.81 ± 0.04	275
Scar	0.36 ± 0.06	0.50 ± 0.08	0.77 ± 0.06	115
Sun damage	0.36 ± 0.09	0.28 ± 0.07	0.92 ± 0.03	78
Mean	0.30 ± 0.01	0.50 ± 0.01	0.79 ± 0.00	171.30

Table B.20
DermX+ localization performance for localizable characteristics (mean±std) with regard to the fuzzy dermatologist attention maps.

	F1 score	Sensitivity	Specificity	Samples
Closed comedo	0.10 ± 0.09	0.11 ± 0.08	0.96 ± 0.02	83
Dermatoglyph disruption	0.60 ± 0.12	0.60 ± 0.15	0.97 ± 0.02	34
Open comedo	0.06 ± 0.05	0.06 ± 0.05	0.97 ± 0.02	93
Papule	0.18 ± 0.04	0.19 ± 0.06	0.97 ± 0.02	232
Patch	0.53 ± 0.05	0.56 ± 0.06	0.89 ± 0.03	191
Plaque	0.61 ± 0.05	0.65 ± 0.06	0.91 ± 0.01	312
Pustule	0.03 ± 0.03	0.04 ± 0.05	0.98 ± 0.01	91
Scale	0.49 ± 0.07	0.58 ± 0.11	0.89 ± 0.03	224
Scar	0.35 ± 0.12	0.35 ± 0.15	0.92 ± 0.04	93
Sun damage	0.58 ± 0.20	0.53 ± 0.19	0.90 ± 0.10	47
Mean	0.36 ± 0.01	0.37 ± 0.01	0.94 ± 0.00	138.00

Appendix A. Additional dataset information

See Figs. A.11–A.13 and Tables A.8–A.11.

Appendix B. Model training and extended performance

See Tables B.12–B.21.

Table B.21
DermX+ characteristics localization performance with regard to fuzzy dermatologist attention maps. The results include the localization performance of characteristics identified by the dermatologists but not by the model.

	F1 score	Sensitivity	Specificity	Samples
Closed comedo	0.09 ± 0.08	0.09 ± 0.07	0.97 ± 0.02	96
Dermatoglyph disruption	0.39 ± 0.17	0.4 ± 0.21	0.98 ± 0.02	54
Open comedo	0.05 ± 0.04	0.05 ± 0.04	0.98 ± 0.02	110
Papule	0.15 ± 0.04	0.16 ± 0.06	0.97 ± 0.02	278
Patch	0.41 ± 0.07	0.43 ± 0.07	0.92 ± 0.02	249
Plaque	0.54 ± 0.06	0.58 ± 0.07	0.92 ± 0.01	352
Pustule	0.03 ± 0.03	0.03 ± 0.04	0.98 ± 0.01	106
Scale	0.41 ± 0.07	0.48 ± 0.09	0.91 ± 0.02	275
Scar	0.29 ± 0.1	0.29 ± 0.13	0.93 ± 0.04	115
Sun damage	0.34 ± 0.14	0.31 ± 0.15	0.94 ± 0.08	78
Mean	0.27 ± 0.01	0.28 ± 0.01	0.95 ± 0.00	171.30

References

Adebayo, J., Gilmer, J., Muelly, M., Goodfellow, I., Hardt, M., Kim, B., 2018. Sanity checks for saliency maps. In: Proceedings of the 32nd International Conference on Neural Information Processing Systems. pp. 9525–9536.

Barata, C., Celebi, M.E., Marques, J.S., 2021. Explainable skin lesion diagnosis using taxonomies. Pattern Recognit. 110, 107413.

Burlina, P.M., Joshi, N.J., Ng, E., Billings, S.D., Rebman, A.W., Aucott, J.N., 2019. Automated detection of erythema migrans and other confounding skin lesions via deep learning. Comput. Biol. Med. 105, 151–156.

Deng, J., Dong, W., Socher, R., Li, L.-J., Li, K., Fei-Fei, L., 2009. Imagenet: A large-scale hierarchical image database. In: 2009 IEEE Conference on Computer Vision and Pattern Recognition. Ieee, pp. 248–255.

DermNetNZ, 2021. https://dermnetnz.org/, Accessed: 2021-04-01.

DeYoung, J., Jain, S., Rajani, N.F., Lehman, E., Xiong, C., Socher, R., Wallace, B.C., 2020. ERASER: A benchmark to evaluate rationalized NLP models. Trans. Assoc. Comput. Linguist..

Esteva, A., Kuprel, B., Novoa, R.A., Ko, J., Swetter, S.M., Blau, H.M., Thrun, S., 2017. Dermatologist-level classification of skin cancer with deep neural networks. Nature 542 (7639), 115–118.

Federman, D.G., Concato, J., Kirsner, R.S., 1999. Comparison of dermatologic diagnoses by primary care practitioners and dermatologists: a review of the literature. Arch. Family Med. 8 (2), 170.

Feng, H., Berk-Krauss, J., Feng, P.W., Stein, J.A., 2018. Comparison of dermatologist density between urban and rural counties in the United States. JAMA Dermatol. 154 (11), 1265–1271.

Goodman, B., Flaxman, S., 2017. European union regulations on algorithmic decision-making and a "right to explanation". AI Mag. 38 (3), 50–57.

Graziani, M., Palatnik de Sousa, I., Vellasco, M.M., Costa da Silva, E., Müller, H., Andrearczyk, V., 2021. Sharpening local interpretable model-agnostic explanations for histopathology: Improved understandability and reliability. In: International Conference on Medical Image Computing and Computer-Assisted Intervention. Springer, pp. 540–549.

Han, S.S., Park, I., Chang, S.E., Lim, W., Kim, M.S., Park, G.H., Chae, J.B., Huh, C.H., Na, J.-I., 2020. Augmented intelligence dermatology: deep neural networks empower medical professionals in diagnosing skin cancer and predicting treatment options for 134 skin disorders. J. Invest. Dermatol. 140 (9), 1753–1761.

Hay, R.J., Johns, N.E., Williams, H.C., Bolliger, I.W., Dellavalle, R.P., Margolis, D.J., Marks, R., Naldi, L., Weinstock, M.A., Wulf, S.K., et al., 2014. The global burden of skin disease in 2010: an analysis of the prevalence and impact of skin conditions. J. Invest. Dermatol. 134 (6), 1527–1534.

He, K., Zhang, X., Ren, S., Sun, J., 2016. Deep residual learning for image recognition. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. pp. 770–778.

Jacovi, A., Goldberg, Y., 2020. Towards faithfully interpretable NLP systems: How should we define and evaluate faithfulness? In: Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics. pp. 4198–4205.

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

Karimkhani, C., Dellavalle, R.P., Coffeng, L.E., Flohr, C., Hay, R.J., Langan, S.M., Nsoesie, E.O., Ferrari, A.J., Erskine, H.E., Silverberg, J.I., Vos, T., Naghavi, M., 2017. Global skin disease morbidity and mortality: An update from the global burden of disease study 2013. JAMA Dermatol. 153 (5), 406–412. http://dx.doi.org/10.1001/jamadermatol.2016.5538.

Kelly, C.J., Karthikesalingam, A., Suleyman, M., Corrado, G., King, D., 2019. Key challenges for delivering clinical impact with artificial intelligence. BMC Med. 17 (1), 1–9.

Kringos, D.S., Boerma, W.G., Hutchinson, A., Saltman, R.B., Organization, W.H., et al., 2015. Building Primary Care in a Changing Europe. World Health Organization. Regional Office for Europe.

Li, K., Wu, Z., Peng, K.-C., Ernst, J., Fu, Y., 2018. Tell me where to look: Guided attention inference network. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. pp. 9215–9223.

Lim, H.W., Collins, S.A., Resneck Jr., J.S., Bolognia, J.L., Hodge, J.A., Rohrer, T.A., Van Beek, M.J., Margolis, D.J., Sober, A.J., Weinstock, M.A., et al., 2017. The burden of skin disease in the United States. J. Am. Acad. Dermatol. 76 (5), 958–972.

Loshchilov, I., Hutter, F., 2016. SGDR: Stochastic gradient descent with warm restarts.

Loshchilov, I., Hutter, F., 2018. Decoupled weight decay regularization. In: International Conference on Learning Representations.

Lundberg, S.M., Lee, S.-I., 2017. A unified approach to interpreting model predictions. In: Guyon, I., Luxburg, U.V., Bengio, S., Wallach, H., Fergus, R., Vishwanathan, S., Garnett, R. (Eds.), Advances in Neural Information Processing Systems. Vol. 30, Curran Associates, Inc..

Mathew, B., Saha, P., Yimam, S.M., Biemann, C., Goyal, P., Mukherjee, A., 2021. Hatexplain: A benchmark dataset for explainable hate speech detection. In: Proceedings of the AAAI Conference on Artificial Intelligence. Vol. 35, (17), pp. 14867–14875.

Nast, A., Griffiths, C.E.M., Hay, R., Sterry, W., Bolognia, J.L., 2016. The 2016 international league of dermatological societies' revised glossary for the description of cutaneous lesions. Br. J. Dermatol. 174 (6), 1351–1358.

Oakley, A., 2017. Dermatology Made Easy. Scion Publishing Ltd, The Old Hayloft, Vantage Business Park, Bloxham Road, Banbury OX16 9UX, UK.

Redmon, J., Farhadi, A., 2018. Yolov3: An incremental improvement. arXiv preprint arXiv:1804.02767.

Ribeiro, M.T., Singh, S., Guestrin, C., 2016. "Why should I trust you?" explaining the predictions of any classifier. In: Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining. pp. 1135–1144.

Selvaraju, R.R., Cogswell, M., Das, A., Vedantam, R., Parikh, D., Batra, D., 2017. Grad-CAM: Visual explanations from deep networks via gradient-based localization. In: Proceedings of the IEEE International Conference on Computer Vision. pp. 618–626.

Sickles, E., D'Orsi, C., Bassett, L., et al., 2013. ACR BI-RADS mammography. In: ACR BI-RADS® Atlas, Breast Imaging Reporting and Data System. American College of Radiology.

Simonyan, K., Vedaldi, A., Zisserman, A., 2014. Deep inside convolutional networks: Visualising image classification models and saliency maps. arXiv:1312.6034.

Singh, A., Sengupta, S., Lakshminarayanan, V., 2020. Explainable deep learning models in medical image analysis. J. Imaging 6 (6), 52.

Sun, X., Yang, J., Sun, M., Wang, K., 2016. A benchmark for automatic visual classification of clinical skin disease images. In: European Conference on Computer Vision. Springer, pp. 206–222.

Tan, M., Le, Q., 2019. EfficientNet: Rethinking model scaling for convolutional neural networks. In: International Conference on Machine Learning. PMLR, pp. 6105–6114.

Tan, M., Pang, R., Le, Q.V., 2020. Efficientdet: Scalable and efficient object detection. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 10781–10790.

Thomsen, K., Iversen, L., Titlestad, T.L., Winther, O., 2020. Systematic review of machine learning for diagnosis and prognosis in dermatology. J. Dermatol. Treatment 31 (5), 496–510.

Topol, E.J., 2019. High-performance medicine: the convergence of human and artificial intelligence. 25, (1), pp. 44–56.

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

Tschandl, P., Rosendahl, C., Kittler, H., 2018. The HAM10000 dataset, a large collection of multi-source dermatoscopic images of common pigmented skin lesions. Sci. Data 5 (1), 1–9.

V7-Labs, 2021. Darwin V7-labs. https://darwin.v7labs.com, Accessed: 2021-05-01.

Zhang, J., Xie, Y., Xia, Y., Shen, C., 2019. Attention residual learning for skin lesion classification. IEEE Trans. Med. Imaging 38 (9), 2092–2103.

APPENDIX B

Dermatological Diagnosis Explainability Benchmark for Convolutional Neural Networks

Unpublished.

Dermatological Diagnosis Explainability Benchmark for Convolutional Neural Networks

A Preprint

Raluca Jalaboi[1,2], Ole Winther[1,3,4], and Alfiia Galimzianova[2]

[1]Department of Applied Mathematics and Computer Science at the Technical University of Denmark, Richard Petersens Plads, Building 324, DK-2800 Kongens Lyngby, Denmark
[2]Medable A/S, Havnegade 25, 3., DK-1058 Copenhagen C, Denmark
[3]Bioinformatics Centre, Department of Biology, University of Copenhagen, Copenhagen, Denmark
[4]Center for Genomic Medicine, Rigshospitalet, Copenhagen University Hospital, Copenhagen, Denmark

Abstract

In recent years, large strides have been taken in developing machine learning methods for various dermatological applications, supported in part by the widespread success of deep learning. To date, diagnosing diseases from images is one of the most explored applications of deep learning within dermatology. Convolutional neural networks (ConvNets) are the most commonly used deep learning method in medical imaging due to their training efficiency and accuracy, although they are often described as black boxes because of their limited explainability. One popular way to obtain insight into a ConvNet's decision mechanism is gradient class activation maps (Grad-CAM). A quantitative evaluation of the Grad-CAM explainability has been recently made possible by the release of DermXDB, a skin disease diagnosis explainability dataset which enables benchmarking the explainability performance of ConvNet architectures. In this paper, we perform a literature review to identify the most common ConvNet architectures used for this task, and compare their Grad-CAM explainability performance with the explanation maps provided by DermXDB. We identified 11 architectures: DenseNet121, EfficientNet-B0, InceptionV3, InceptionResNetV2, MobileNet, MobileNetV2, NASNetMobile, ResNet50, ResNet50V2, VGG16, and Xception. We pre-trained all architectures on an clinical skin disease dataset, and then fine-tuned them on a subset of DermXDB. Validation results on the DermXDB holdout subset show an explainability F1 score of between 0.35-0.46, with Xception the highest explainability performance, while InceptionResNetV2, ResNet50, and VGG16 displaying the lowest. NASNetMobile reports the highest characteristic-level explainability sensitivity, despite it's mediocre diagnosis performance. These results highlight the importance of choosing the right architecture for the desired application and target market, underline need for additional explainability datasets, and further confirm the need for explainability benchmarking that relies on quantitative analyses rather than qualitative assessments.

Keywords deep learning, dermatologys, explainability, benchmark, review

1 Introduction

With an expected shortage of approximately ten million healthcare professionals by 2030 [World Health Organization, 2016], the world is facing a massive healthcare crisis. Automation has been proposed as a solution to the scarcity of medical professionals, with the Food and Drugs Administration in the United States approving medical devices based on artificial intelligence for marketing to the public [U.S. Food and Drug Administration, 2018].

This development is due in part to the advancement in machine learning using unstructured data. Ever since Krizhevsky et al. [2017] won the ImageNet Large Scale Visual Recognition Challenge [Russakovsky et al., 2015] using a convolutional neural network (ConvNet), ConvNets have been at the forefront of machine learning based automation. Employed primarily in healthcare for imaging applications, ConvNets have been used for disease diagnosis [Gao et al., 2019], cell counting [Falk et al., 2019], disease severity assessment [Gulshan et al., 2016], disease progression estimation [Kijowski et al., 2020], lesion or anatomical region segmentation [Hesamian et al., 2019, Ramesh et al., 2021], etc. Esteva et al. [2017] were the first to demonstrate that ConvNets can achieve expert-level performance in dermatological diagnosis using dermoscopy images. Since then, dermatology has embraced ConvNets as a solution to various diagnosis and segmentation tasks [Esteva et al., 2017, Zhang et al., 2019, Jinnai et al., 2020, Haenssle et al., 2020, Roy et al., 2022].

Despite these considerable advancements in medical imaging, there has not yet been a widespread adoption of machine learning based automation in the clinical workflow. One of the main hurdles that detract from adoption is the lack of ConvNet explainability [Kelly et al., 2019], this issue being enhanced by the recently implemented legislation aimed at ensuring that automated methods can offer an explanation into their decision mechanisms [Goodman and Flaxman, 2017]. Different post-hoc explainability methods have been proposed as a way to explain a ConvNet's decisions [Bai et al., 2021, Selvaraju et al., 2017, Lundberg and Lee, 2017, Ribeiro et al., 2016]. Gradient class activation maps (Grad-CAM) is currently the most commonly used explainability method within medical imaging, due to its intrinsic ease of interpretation and its low computational requirements. However, validating the resulting explanations is an expensive, time consuming process that requires domain expert intervention, and thus most explainability validations are performed as small, qualitative analyses. With the release of DermXDB [Jalaboi et al., 2022], it became possible to quantitatively analyse the explainability of ConvNets trained for diagnosing six skin conditions: acne, psoriasis, seborrheic dermatitis, viral warts, and vitiligo.

The purpose of this benchmark is to provide the means to quantitatively compare the explainability of the state-of-the-art approaches to dermatological diagnosis using photographic imaging. Our contributions are twofold:

1. We perform a comprehensive systematic review to reveal the usage of the ConvNets for the task of dermatological diagnosis using photographic images,
2. We benchmark the identified ConvNets for diagnostic and explainability performance and compare them with eight expert dermatologists.

2 Background

2.1 Machine learning methods in dermatological diagnosis

After the renewed interest in artificial intelligence and machine learning that started in 2012, practitioners from both academia and the industry began investigating automated methods for dermatological applications [Thomsen et al., 2020, Jeong et al., 2022]. Until 2017, the vast majority of articles applying machine learning methods on dermatological problems were using classical models such as support vector machines [Liu et al., 2012, Sabouri et al., 2014], and linear or logistic regression [Kaur et al., 2015, Kefel et al., 2016]. These models were trained using hand-crafted features or features extracted using classical computer vision methods such as gray-level co-occurrence matrices [Shimizu et al., 2014], Sobel and Hessian filters [Arroyo and Zapirain, 2014], or HOS texture extraction [Shrivastava et al., 2016]. However, the main drawback of classical computer vision approaches is that hand-crafting features is an expensive, time-consuming process, while their automated extraction is too sensitive to the environmental factors of the image acquisition (e.g. lighting, zoom).

Esteva et al. [2017] were the first to propose a ConvNet for diagnosing skin conditions from dermoscopy images. Their ConvNet reached expert-level performance without requiring any hand-crafted features or classical computer vision models, thus paving the way towards the current popularity of ConvNets in dermatological applications.

One key component to the rise of ConvNets was the introduction of large scale dermatological datasets. The International Skin Imaging Collaboration (ISIC) challenge dataset [Codella et al., 2018] is one of the best known open access dermoscopy datasets, containing 25,331 images distributed over nine diagnostic categories. Large clinical image datasets are also available for research purposes, such as SD-260 [Sun et al., 2016] which consists of 20,600 clinical images of 260 different skin diseases, and DermNetNZ [DermNetNZ, 2021] which contains more than 25,000 clinical images.

Aided by the release of increasingly more performant architectures, their publicly available pre-trained weights on the ImageNet [Deng et al., 2009] dataset, and the recently published public dermatological datasets, the vast majority of research contributions in machine learning applications for dermatology rely on ConvNet architectures. ConvNets have been extensively used in lesion diagnosis [Tschandl et al., 2017, Han et al., 2018, Reshma et al., 2022] and lesion segmentation [Yuan et al., 2017, Wu et al., 2022, Baig et al., 2020] on different modalities relevant for the domain. Attempts at explaining the decisions taken by ConvNets were made by several groups [Tschandl et al., 2020, Tanaka et al., 2021], but no quantitative analysis was performed.

Table 1: Search query used on PubMed to identify the list of relevant articles. We searched for articles focused on dermatology, using deep learning methods, written in English. The query was last performed on the 20th of February 2023.

Search term		Search term		Search term
(((dermatology[MeSH Terms]) OR (skin disease[MeSH Terms]) OR (skin lesion[MeSH Terms]))	AND	((neural network[MeSH Terms]) OR (machine learning[MeSH Terms]) OR (artificial intelligence[MeSH Terms]) OR (deep learning) OR (deep neural network) OR (convolutional neural network))	AND	(English[Language])

2.2 Explainability in convolutional neural networks

ConvNets have, from their very beginning, been notoriously difficult to interpret and explain. Interpretability is generally considered the ability to understand the internal structure and properties of a ConvNet architecture, while explainability is defined as a ConvNet's capacity to offer plausible arguments in favour of its decision [Roscher et al., 2020]. Within healthcare, explainability is especially important due to its intrinsic ability to interact with domain experts in a common vocabulary [Kelly et al., 2019]. Although some architecture or domain-specific explainability methods exist, most medical imaging research articles employ attribution-based methods due to their ease of use and open source access [Singh et al., 2020, Bai et al., 2021].

There are two main ways of implementing attribution-based methods: through perturbation and by using the ConvNet's gradients. Perturbation-based methods, such as Shapley values [Lipovetsky and Conklin, 2001], LIME [Ribeiro et al., 2016], or SharpLIME [Graziani et al., 2021], rely on modifying the original image and then evaluating the changes in the ConvNet's prediction. For example, LIME uses a superpixel algorithm to split the image into sections, and randomly selects a subset of superpixels to occlude. The target ConvNet then performs an inference step on the perturbed image. This procedure is run multiple times to identify the superpixels that lead to the most drastic change in the ConvNet's prediction. SharpLIME uses hand-crafted segmentations to split the image into relevant sections, and then proceeds with the perturbation process defined in LIME. The main drawback of perturbation based methods is the need to run the prediction algorithm multiple times, which leads to high computational costs and long running times.

Gradient-based methods, such as saliency maps [Simonyan and Zisserman, 2015], guided backpropagation [Springenberg et al., 2014], gradient class-activation maps (Grad-CAM) [Selvaraju et al., 2017], or layer-wise relevance propagation [Bach et al., 2015], use a ConvNet's backpropagation step to identify the areas in an image that contribute the most to the prediction. In general, gradient-based methods compute the gradient of a given input in relation to the prediction, and apply different post-processing methods to the output. In the case of Grad-CAM, image features are extracted by forward propagating the image until the last convolutional layer. Then, the gradient is set to 0 for all classes except the target class, and the signal is backpropagated to the last convolutional layer. The extracted image features that directly contribute to the backpropagated signal constitute the Grad-CAM for the given class. Since the analysis can be performed at the same time as the inference itself and only requires one iteration, Grad-CAM is often used in research and industrial applications [Pereira et al., 2018, Young et al., 2019, Tschandl et al., 2020, Hepp et al., 2021, Jalaboi et al., 2023]. Due to its popularity, in this paper we will use Grad-CAM to benchmark the explainability of commonly used ConvNet architectures.

3 Material and methods

3.1 Literature review

We performed a systematic literature review on PubMed, following the methodology introduced by Thomsen et al. [2020]. The query, described in Table 1, focused on dermatological applications of deep learning. A total of 3,650 articles were retrieved. We excluded articles that focused on domains other than dermatology, articles that did not include an original contribution in disease classification, articles using modalities other than photographic images, articles using methods other than ConvNets, and articles using proprietary ConvNets.

3.2 Explainability benchmark

3.2.1 Explainability dataset

For explainability benchmarking, we use DermXDB, a skin disease diagnosis explainability dataset published by Jalaboi et al. [2022]. The dataset consists of 524 images sourced from DermNetNZ [DermNetNZ, 2021] and SD-260 [Sun et al., 2016], and labeled with diagnoses and explanations in the form of visual skin lesion characteristics by eight board-certified dermatologists. To match the Grad-CAM output, we focus on the characteristic localization task.

3.2.2 Diagnosis evaluation

For establishing the expert-level diagnosis performance, we compare each dermatologist with the reference standard diagnosis. We follow the same approach for benchmarking the diagnosis performance of the ConvNets. We evaluate the performance using the categorical F1 score, sensitivity, and specificity, defined as:

$$\text{F1 score} = \frac{2TP}{2TP + FP + FN}, \tag{1}$$

$$\text{Sensitivity} = \frac{TP}{TP + FN}, \tag{2}$$

$$\text{Specificity} = \frac{TN}{TN + FP}, \tag{3}$$

where the true positives TP represent correctly classified samples, the false positives FP represent samples incorrectly classified as part of the target class, the false negatives FN represent samples of the target class incorrectly classified as being part of a different class, and the true negatives TN represent samples correctly identified as not being part of the target class.

3.2.3 Explainability evaluation

For establishing expert-level explainability performance, we compare the attention masks of each dermatologist with the aggregated fuzzy union of attention masks created by the other seven dermatologists (explanation maps). More specifically, we define the *image-level explanation maps* as the union of all characteristics segmented by all dermatologists for an image, and the *characteristic-level explanation maps* as the union of all segmentations for each characteristic for an image. Figure 1 illustrates the mask creation process for a psoriasis case. The ConvNet Grad-CAM attention maps are compared with explanations maps derived from all eight dermatologist evaluations.

These two types of explanation maps offer a way to check whether the ConvNets take into account the entire area selected by dermatologists as important to their decision, and whether they focus on specific characteristics when making their decisions. To quantify the similarity between the Grad-CAMs and the explanation maps, we compute the F1 score, sensitivity and specificity following their fuzzy implementation defined in [Crum et al., 2006], described as:

$$\text{F1 score} = \frac{2\sum_{p \in pixels} \min(\mathcal{G}_p, \mathcal{E}_p)}{\sum_{p \in pixels}(\mathcal{G}_p) + \sum_{p \in pixels}(\mathcal{E}_p)}, \tag{4}$$

$$\text{Sensitivity} = \frac{\sum_{p \in pixels} \min(\mathcal{G}_p, \mathcal{E}_p)}{\sum_{p \in pixels}(\mathcal{S}_p)}, \tag{5}$$

$$\text{Specificity} = \frac{\sum_{p \in pixels} \min(1 - \mathcal{G}_p, 1 - \mathcal{E}_p)}{\sum_{p \in pixels}(1 - \mathcal{E}_p)}, \tag{6}$$

where $\mathcal{G}$ is the ConvNet-generated Grad-CAM, and $\mathcal{E}$ is the explanation map for a given image.

For characteristics, we report the Grad-CAM sensitivity with regard to the characteristic-level explanation maps. Specificity and F1 score were considered too stringent, as multiple characteristics can be present and essential for a diagnosis, and an explainable ConvNet must detect all of them to plausibly explain the diagnosis.

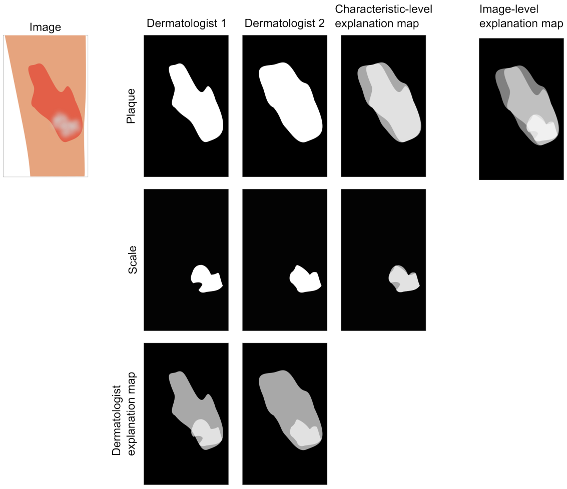


Image
Dermatologist 1
Dermatologist 2
Characteristic-level explanation map
Image-level explanation map
Plaque
Scale
Dermatologist explanation map

Figure 1: Explanation maps creation example for a psoriasis case evaluated by two dermatologists. Both dermatologists identified plaque and scale as the two characteristics associated with the psoriasis diagnosis, and localized them. By combining the localization maps for each characteristic, we obtain the characteristic-level explanation maps. By combining the localization maps created by each dermatologist, we obtain the individual dermatologist explanation maps. By combining all localization maps, we obtain the image-level explanation map.

3.2.4 Experimental setup

From the 22 articles that fulfilled all inclusion criteria, we selected the set of ConvNets to benchmark based on their reproducibility: we required that all benchmarked ConvNets had been pre-trained on ImageNet due to the limited amount of training data available. Thus, we exclude architectures that do not have publicly available pre-trained ImageNet weights compatible with the deep learning Keras framework [Chollet, 2015], i.e. GoogLeNet [Szegedy et al., 2015], InceptionV4 [Babenko and Lempitsky, 2015], MobileNetV3 [Howard et al., 2019], SENet [Hu et al., 2018], SE-ResNet [Hu et al., 2018], SEResNeXT [Hu et al., 2018], and ShuffleNet [Zhang et al., 2018]. Furthermore, as several articles compare different versions of the same architecture (e.g. EfficientNet-B0 through EfficientNet-B7, see Table 2), we select the smallest version of each architecture for our benchmark to avoid overfitting to the DermXDB dataset.

In the rest of this work, we will focus on the following ConvNets: DenseNet121 [Huang et al., 2017], EfficientNet-B0 [Tan and Le, 2019], InceptionResNetV2 [Szegedy et al., 2017], InceptionV3 [Szegedy et al., 2016], MobileNet [Howard et al., 2017], MobileNetV2 [Sandler et al., 2018], NASNetMobile [Zoph et al., 2018], ResNet50 [He et al., 2016a], ResNet50V2 [He et al., 2016b], VGG16 [Simonyan and Zisserman, 2015], and Xception [Chollet, 2017].

We used the pre-trained weights offered by Keras to initialize the networks in our experiments. Next, all ConvNets were pre-trained on a proprietary clinical photography skin disease dataset collected by a dermatologist between 2004-2018. All images included in the dataset were anonymized, and the patients consented to their data being used for research purposes. More information about the dataset is available in Appendix Table A1. We performed a hyper-parameter search for each ConvNet, with the values used for experimentation and the validation performance being reported in Appendix Table A2 and Appendix Table A3, respectively. We further fine-tuned all ConvNets for 50 epochs with 261 randomly chosen images from the DermXDB dataset. The remaining 263 images were used as the test set. Each ConvNet was trained and tested five times. All results presented in this paper are aggregated over the five test runs. All code used for running the experiments is available at `https://github.com/ralucaj/dermx-benchmark`.

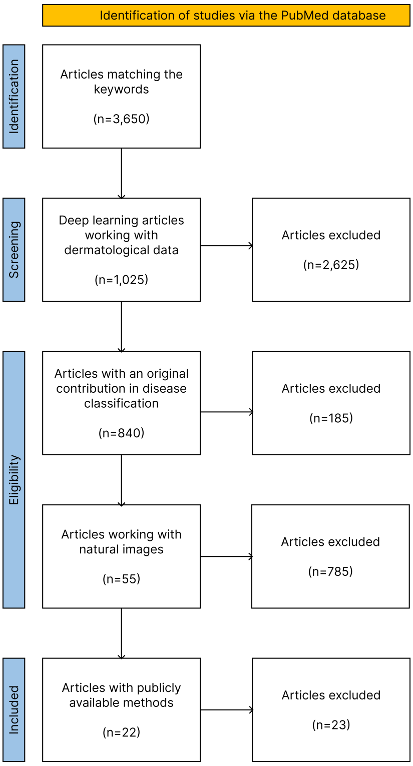

Identification of studies via the PubMed database
Identification
Screening
Eligibility
Included
Articles matching the keywords (n=3,650)
Deep learning articles working with dermatological data (n=1,025)
Articles excluded (n=2,625)
Articles with an original contribution in disease classification (n=840)
Articles excluded (n=185)
Articles working with natural images (n=55)
Articles excluded (n=785)
Articles with publicly available methods (n=22)
Articles excluded (n=23)

Figure 2: The Preferred Reporting Items for Systematic Reviews and Meta Analyses (PRISMA) statement flowchart of the performed review process for identifying the benchmarked ConvNet architectures. First, we screened articles to ensure that they were using dermatological data and deep learning methods. Afterwards, we excluded review articles and contributions focused on tasks other than classification, and articles that that used non-photographic image data, e.g. dermoscopy, whole slides. Finally, we excluded articles that used proprietary ConvNets, leading to 22 articles serving as the benchmark basis.

4 Results

4.1 Literature review

Figure 2 displays the Preferred Reporting Items for Systematic Review and Meta-Analyses statement flowchart of the performed review, while Figure 3 illustrates the evolution of articles topics over the years. Out of the original 3,650 articles, only 22 fulfilled all the inclusion criteria. Table 2 summarizes the ConvNet architectures, their implementation, and reported performance employed in the final 22 articles selected for benchmarking.

Table 2: Overview of the 22 articles fulfilling all inclusion criteria. All articles use ConvNets for a dermatological classification task using photographic images. Tasks vary between binary or multi-disease diagnosis, disease risk assessment, lesion type classification, and severity assessment.

Publication	ConvNets employed	Task	Data	Performance
Aggarwal [2019]	InceptionV3	Disease diagnosis on five classes	Open source images and images scraped from Google	0.66 F1 score, 0.65 sensitivity, 0.91 specificity, 0.67 precision, 0.91 NPV, 0.57 MCC
Burlina et al. [2019]	ResNet50	Disease diagnosis on four classes	Internet-scraped images	82.79% accuracy, 0.76 kappa score
Zhao et al. [2019]	Xception	Skin cancer risk assessment with three classes	Clinical images	72% accuracy, 0.92-0.96 ROC AUC, 0.85-0.93 sensitivity, 0.85-0.91 specificity
Burlina et al. [2020]	ResNet50, ResNet152, InceptionV3, InceptionResNetV2, DenseNet	Disease diagnosis on eight classes	Clinical and other photographic images scraped using Google and Bing	71.58% accuracy, 0.70 sensitivity, 0.96 specificity, 0.72 precision, 0.96 NPV, 0.67 kappa, 0.72 F1 score, 0.80 average precision, 0.94 AUC
Chin et al. [2020]	DenseNet121, VGG16, ResNet50	Binary skin cancer risk assessment	Smartphone images	0.83-0.86 AUC, 0.72-0.77 sensitivity, 0.85-0.86 specificity
Han et al. [2020]	SENet, SE-ResNet50, VGG19	Disease classification on 134 classes	Clinical images	44.8-56.7% accuracy, 0.94-0.98 AUC
Liu et al. [2020]	InceptionV4	Disease diagnosis on 26 classes	Clinical images	66% accuracy, 0.56 sensitivity
Zhao et al. [2020]	DenseNet121, Xception, InceptionV3, InceptionResNetV2	Binary psoriasis classification	Clinical images	96% accuracy, 0.95-0.98 AUC, 0.96-0.97 specificity, 0.83-0.95 sensitivity
Wu et al. [2021]	SEResNeXt, SE-ResNet, InceptionV3	Disease diagnosis on five classes	Clinical images	0.96-0.97 AUC, 90-91% accuracy, 0.90-0.93 sensitivity, 0.90 specificity
Aggarwal and Papay [2022]	InceptionResNetV2	Disease diagnosis on four classes	Clinical images	0.60-0.82 sensitivity, 0.60-0.82 specificity, 0.33-0.93 precision, 0.33-0.93 NPV, 0.43-0.84 F1 score
Ba et al. [2022]	EfficientNet-B3	Disease diagnosis on 10 classes	Clinical images	78.45% accuracy, 0.73 kappa
Hossain et al. [2022]	VGG16, VGG19, ResNet50, ResNet101, ResNet50V2, ResNet101V2, InceptionV3, InceptionV4, InceptionResNetV2, Xception, DenseNet121, DenseNet169, DenseNet201, MobileNetV2, MobileNetV3Small, MobileNetV3Large, NASNetMobile, EfficientNet-B0 through EfficientNet-B5	Binary Lyme disease classification	Smartphone images	61.42-84.42% accuracy, 0.72-0.90 sensitivity, 0.50-0.81 specificity, 0.61-0.83 precision, 0.63-0.87 NPV, 0.23-0.69 MCC, 0.22-0.69 Cohen's kappa, 1.46-4.70 positive likelihood ratio, 0.14-0.55 negative likelihood ratio, 0.66-0.0.85 F1 score, 0.65-0.92 AUC

Hüsers et al. [2022]	MobileNet	Binary wound maceration classification	Clinical images	69% accuracy, 0.69 sensitivity, 0.67 precision
Liu et al. [2022]	InceptionResNetV2	Ulcer characteristic diagnosis on two and three classes	Clinical images	71.2-99.4% accuracy, 0.68-0.99 sensitivity, 0.71-1.00 precision, 0.70-0.94 F1 score
Malihi et al. [2022]	Xception	Binary wound type classification	Clinical images	67-83% accuracy, 0.65-0.94 sensitivity, 0.70-0.75 specificity, 0.65-0.75 precision, 0.70-0.85 F1 score
Munthuli et al. [2022]	DenseNet121	Skin lesion severity classification with five classes	Smartphone images	0.43-0.91 sensitivity, 0.80-0.98 specificity, 0.50-0.87 F1 score
Ni et al. [2022]	DenseNet121, ResNet50	Radiation dermatitis severity classification on four classes	Clinical images	83% accuracy, 0.74-1.00 F1 score
Roy et al. [2022]	ResNet101	Disease diagnosis on 26 classes	Clinical images	62.6% - 75.6% accuracy, 69.3-81.8 AUPR
Sahin et al. [2022]	ResNet18, GoogleNet, EfficientNet-B0, NASNetMobile, ShuffleNet, MobileNetV2	Binary monkeypox classification	Smartphone images	73.33-91.11% accuracy
Xia et al. [2022]	ResNet50	Binary skin cancer classification	Smartphone images	0.77-0.82 AUC, 0.76-0.79 AP
Zhou et al. [2022]	ResNet50	Disease diagnosis on three classes	Clinical images	0.32 error rate, 0.68 sensitivity, 0.69 precision, 0.68 F1 score
Zhang and Ma [2022]	ResNet50	Acne severity classification with three classes	Clinical images	74% accuracy

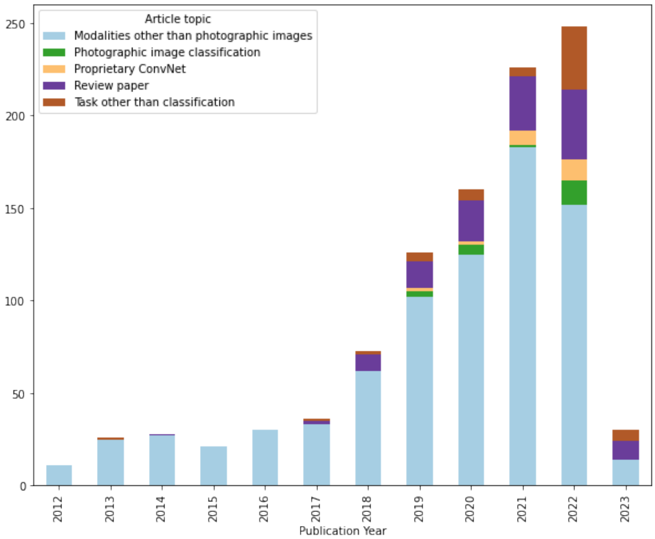


Figure 3: Distribution of retrieved article topics per publication year, based on the search query defined in Table 1 (ran on the 20th of February 2023). 2017 marks an explosion in the number of deep learning applications in dermatology, a fact highlighted by the large increase in articles in the subsequent years, and an increase in review articles. Starting 2019, the industrial involvement in this field became apparent due to the increase in proprietary ConvNets. 2019 also marks the first emergence of dermatological applications using photographic imaging. Finally, although classification is still the most common application, other applications are becoming increasingly more researched.

4.2 Diagnosis results

Table 3 provides an overview of the diagnostic performance of the networks and that of the dermatologists on average in terms of F1 score. As can be seen from the table, although several ConvNets achieve expert-level performance when diagnosing actinic keratosis, seborrheic dermatitis, and viral warts, none of them achieve overall expert-level performance. ConvNets follow the trend also seen in dermatologists of having difficulties correctly diagnosing actinic keratosis and seborrheic dermatitis, while the diagnosis of acne and viral warts displays higher performance. Similar trends can be observed for the sensitivity and specificity performance, as seen in Appendix Table A4 and Appendix Table A5, respectively.

4.3 Explainability results

Table 4 shows the image-level explainability results for each of the benchmarked ConvNets, while Figure 4 shows the relationship between ConvNet diagnosis performance, image-level explainability, and number of parameters. Xception scores the highest on the image-level Grad-CAM F1 score, while InceptionResNetV2, ResNet50, and VGG16 have the lowest performance. DenseNet121 and NASNetMobile report expert-level sensitivity scores, while ResNet50V2 achieves expert-level performance in specificity.

Looking at the characteristic-level sensitivity depicted in Figure 5, NASNetMobile and DenseNet121 achieve the highest overall performance. InceptionResNetV2, ResNet50, ResNet50V2, and VGG16 report the lowest scores. All ConvNets outperform dermatologists in closed comedo, open comedo, and pustule. The opposite is true for dermatoglyph disruption, leukotrichia, patch, plaque, scale, sun damage, and telangiectasia – no ConvNet reaches expert-level.

Figure 6 illustrates the differences in Grad-CAMs between the benchmarked ConvNets. Older ConvNet architectures, such as VGG16, InceptionResNetV2, ResNet50, and ResNet50V2, tend to focus on small areas that contain characteristics relevant for the diagnosis, e.g. focusing on a single plaque in the psoriasis diagnosis example, while more modern ConvNets pay attention to the entire area covered by diagnosis-relevant lesions. Several ConvNets, namely

Table 3: Diagnostic performance of the ConvNets (average ± standard deviation across five runs) and dermatologists (average ± standard deviation across eight experts) using F1-score, split by diagnosis. Several ConvNets achieve expert-level per-disease diagnosis performance, in actinic keratosis, seborrheic dermatitis, and viral warts (in **bold**), although none reach the same performance for acne, psoriasis, and vitiligo.

	Acne	**Actinic keratosis**	**Psoriasis**	**Seborrheic dermatitis**	**Viral warts**	**Vitiligo**
ConvNets						
DenseNet121	0.80 ± 0.02	**0.63 ± 0.08**	0.66 ± 0.01	**0.69 ± 0.03**	**0.88 ± 0.03**	0.74 ± 0.03
EfficientNet-B0	0.72 ± 0.03	0.53 ± 0.10	0.60 ± 0.06	**0.57 ± 0.08**	0.80 ± 0.07	0.66 ± 0.02
InceptionV3	0.77 ± 0.02	**0.57 ± 0.11**	0.60 ± 0.02	0.54 ± 0.03	0.77 ± 0.04	0.73 ± 0.05
InceptionResNetV2	0.73 ± 0.02	0.52 ± 0.10	0.53 ± 0.05	0.56 ± 0.05	0.69 ± 0.03	0.53 ± 0.12
MobileNet	0.72 ± 0.06	**0.55 ± 0.19**	0.51 ± 0.14	**0.57 ± 0.06**	0.68 ± 0.06	0.56 ± 0.10
MobileNetV2	0.56 ± 0.07	0.23 ± 0.09	0.31 ± 0.08	0.46 ± 0.05	0.63 ± 0.07	0.48 ± 0.14
NASNetMobile	0.50 ± 0.05	0.33 ± 0.12	0.42 ± 0.07	0.43 ± 0.05	0.55 ± 0.11	0.51 ± 0.05
ResNet50	0.77 ± 0.04	**0.53 ± 0.17**	0.61 ± 0.03	**0.61 ± 0.19**	0.79 ± 0.02	0.61 ± 0.07
ResNet50V2	0.76 ± 0.04	**0.62 ± 0.07**	0.59 ± 0.01	0.57 ± 0.01	0.76 ± 0.01	0.75 ± 0.05
VGG16	0.70 ± 0.05	**0.62 ± 0.03**	0.59 ± 0.03	**0.49 ± 0.15**	0.71 ± 0.03	0.62 ± 0.07
Xception	0.80 ± 0.04	**0.64 ± 0.07**	0.70 ± 0.02	**0.60 ± 0.03**	0.81 ± 0.04	0.81 ± 0.05
Dermatologists						
Average	0.95 ± 0.02	0.79 ± 0.14	0.85 ± 0.06	0.72 ± 0.09	0.93 ± 0.05	0.96 ± 0.03

EfficientNet-B0, MobileNet, MobileNetV2, and VGG16 seem to have overfit on the training set, focusing on the watermark rather than the image itself when diagnosing the vitiligo case.

5 Discussion

5.1 Literature review

ConvNets have become a default approach when it comes to automated diagnosis using images, aligned with the rise of the deep learning methodology for vision recognition. The continuous breakthroughs in diagnostic performance across a wide variety of medical imaging modalities and disorders have made automated diagnosis as close to integration with practice as ever. In dermatology, the diagnosis performance has achieved that of the expert raters as early as 2017 with a seminal work of Esteva et al. [2017] that disrupted the research field and set the trend that still persists, as can be seen through the trends of the continuous growth outlined in Figure 3. The increased interest of industrial entities that started in 2019, illustrated in Figure 3 by the increase in proprietary methods, is further highlighted by the large number of dermatology-oriented med-tech companies relying on machine learning for their products. Year 2019 also marks the year when research groups began investigating photographic images as a primary modality for diagnosing skin conditions, meaning the rise of machine learning solutions to assist teledermatology.

The potential of using ConvNets to streamline dermatological tasks is underlined by the diversity of tasks being solved in the retrieved articles. Classification was the first methodology to be approached, with applications in disease diagnosis, risk assessment, lesion type classification, lesion characteristics identification, and disease severity assessment. Segmentation and natural language processing applications are also gaining more traction, as shown by the constant increase in non-classification tasks in Figure 3.

However, this potential has not yet translated into the much-needed transformation of the clinical practice. In part, this is due to regulatory challenges which are often faced due to the limited generalizability and lack of explainability of the methods [Kelly et al., 2019]. By benchmarking the diagnosis and explainability performance of ConvNets, we both

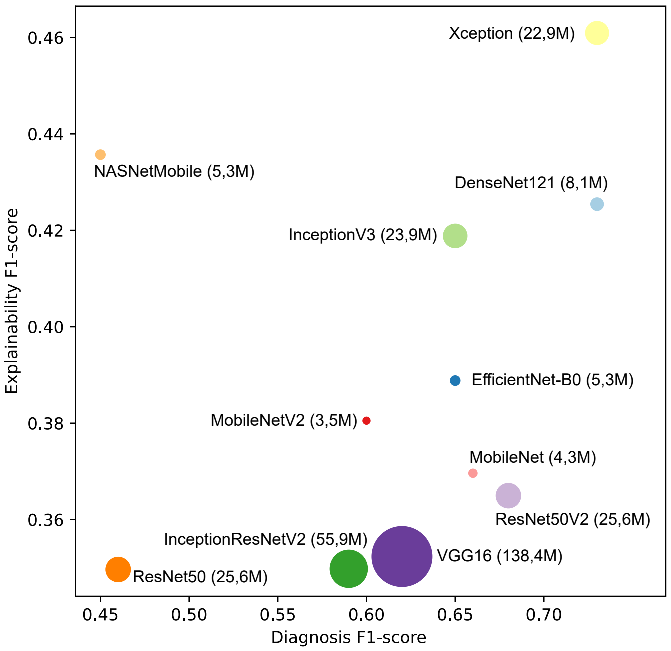


Figure 4: ConvNet explainability as a function of ConvNet performance and their number of parameters. Xception displays both the highest performance and image-level explainability, while ResNet50 performs poorly in both criteria.

enable a comparison among the methods, as well as help the identification gaps between the current state-of-the-art and the clinical practice.

5.2 Diagnosis benchmark

The direct comparison of the diagnostic performance is not possible using reported values from the literature not only due to variability in the choice of the metrics, but more importantly due to the variance in the number of classes and the differences in the datasets used for training and validation (Table 2). By reformulating the task to the diagnosis of six disease classes, utilizing the same initialization, pre-training, and hyperparameter optimization search strategy, and training and validating on the common database, this benchmark minimizes the performance variability related to such implementation details.

We found considerable variability among the diagnostic performance values, with the average F1 scores ranging from 0.50 to 0.80 for acne, from 0.23 to 0.64 for actinic keratosis, from 0.31 to 0.70 for psoriasis, from 0.43 to 0.69 for seborrheic dermatitis, from 0.55 to 0.88 for viral warts, and from 0.51 to 0.81 for vitiligo. These values were aligned with the diagnostic complexity of the diseases as expressed by the performance of the dermatologists, averaging 0.95 for acne, 0.79 for actinic keratosis, 0.85 for psoriasis, 0.72 for seborrheic dermatitis, 0.93 for viral warts, and 0.96 for vitiligo. As such, none of the ConvNets achieved the average dermatologist performance, although there were multiple instances of ConvNets reaching the range of the expert performance for a specific disease (see Table 3). The majority of the benchmarked ConvNets achieved expert level for diagnosis of actinic keratosis and seborrheic dermatitis: seven and six out of 11, respectively. This further confirms the similarity of ConvNet performance with respect to the dermatologists: most ConvNets display a similar difficulty in diagnosing actinic keratosis and seborrheic dermatitis as the eight dermatologists, and a similar ease of diagnosing acne and viral warts.

5.3 Explainability benchmark

While diagnostic performance is recognized as critical for the generalizability of ConvNets, the explainability performance validation has been generally approached as an optional, qualitative, post-hoc analysis. One of the key challenges

Table 4: Explainability performance in terms of the image-level Grad-CAM evaluation for the ConvNets (average ± standard deviation across five runs), and an explanation map evaluation dermatologists (average ± standard deviation across eight experts). Older ConvNets, such as ResNet50, ResNet50V2, and VGG16, have lower performance than most other modern ConvNets. Two networks achieve expert-level sensitivity scores, and one achieves expert-level specificity (in **bold**).

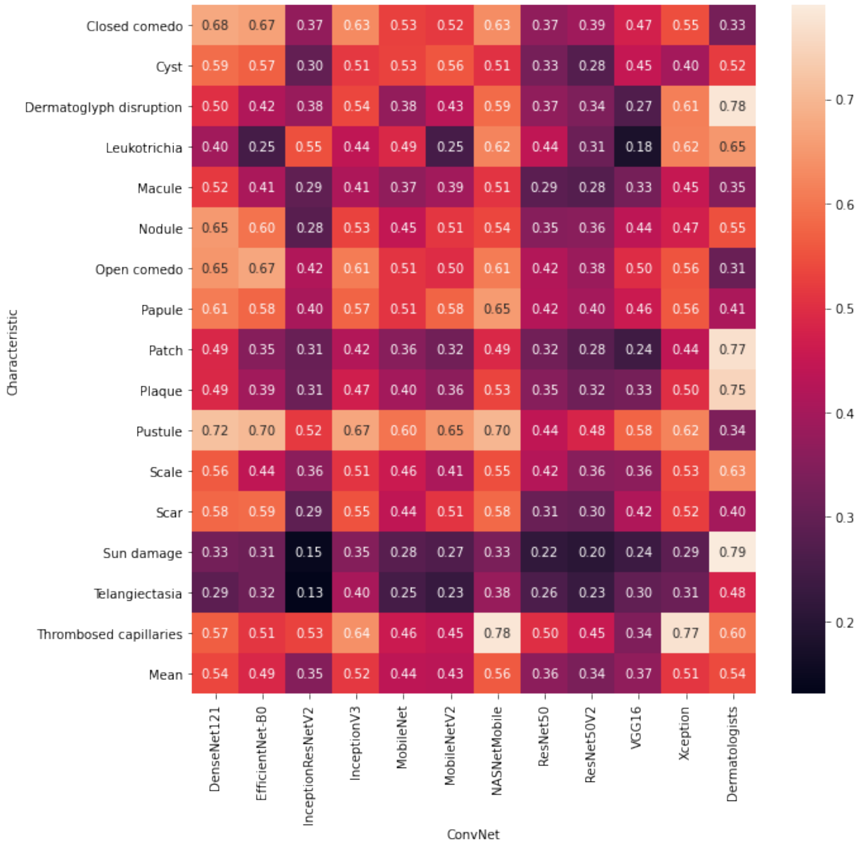
	F1 score	**Sensitivity**	**Specificity**
ConvNets			
DenseNet121	0.43 ± 0.01	$\mathbf{0.61 \pm 0.01}$	0.78 ± 0.00
EfficientNet-B0	0.39 ± 0.01	0.52 ± 0.00	0.82 ± 0.00
InceptionV3	0.42 ± 0.01	0.56 ± 0.01	0.82 ± 0.01
InceptionResNetV2	0.35 ± 0.01	0.40 ± 0.01	0.87 ± 0.01
MobileNet	0.37 ± 0.02	0.50 ± 0.01	0.85 ± 0.01
MobileNetV2	0.38 ± 0.02	0.49 ± 0.02	0.87 ± 0.01
NASNetMobile	0.44 ± 0.00	$\mathbf{0.62 \pm 0.00}$	0.81 ± 0.00
ResNet50	0.35 ± 0.01	0.42 ± 0.03	0.84 ± 0.01
ResNet50V2	0.37 ± 0.01	0.38 ± 0.01	$\mathbf{0.91 \pm 0.00}$
VGG16	0.35 ± 0.01	0.40 ± 0.01	0.86 ± 0.01
Xception	0.46 ± 0.01	0.56 ± 0.00	0.88 ± 0.01
Dermatologists			
Average	0.66 ± 0.03	0.67 ± 0.07	0.93 ± 0.03

faced by researchers trying to implement a more objective validation of explainability is linking the human-approachable explanations with those feasible for ConvNets. With the use of the labels for dermatological diagnosis explainability available from the recently released DermXDB dataset, our benchmark is quantitative as well as predefined. Thus, we avoid potential biases and limitations stemming from machine learning experts with little domain knowledge performing a visual, qualitative evaluation of Grad-CAMs [Tschandl et al., 2020].

The image-level explainability analysis shows that no ConvNet reaches the same F1 score as the dermatologists, although several ConvNets achieve expert-level sensitivity or specificity. Different ConvNets show different patterns of explanation behaviour (Figure 6): some tend to focus on smaller areas that are highly indicative of the target diagnosis, while others tend to focus on the entire affected area. Extensive user tests with both experts and patients would enable us to learn which of the two options is preferred as an explanation: a single, classical lesion descriptive of the diagnosis, or highlighting the entire affected area.

From a characteristic-level sensitivity perspective, most ConvNets outperform the average dermatologist performance in characteristics smaller than 1cm in diameter [Nast et al., 2016]. For larger characteristics, although NASNetMobile and Xception approach expert-level, no ConvNet exceeds it. The relationship between diseases and their characteristics is visible in the characteristic-level ConvNet explainability: most ConvNets report high sensitivity on characteristics often associated with acne and viral warts (e.g. closed and open comedones, papules, and thrombosed capillaries), while reporting a lower performance on characteristics associated with actinic keratosis and seborrheic dermatitis (e.g. plaque, sun damage, and patch). Characteristic-level explainability may be more relevant for use cases where identifying the differentiating factor between different diseases is the most important component for garnering trust.

These result suggests that while ConvNets have the potential to produce human-approachable explanations, more work is necessary to fully achieve expert-level performance. Part of the necessary work is the creation of additional user-derived explainability datasets that enable quantitative analyses on a ConvNet's explainability within a domain. A component of this is performing extensive user tests to identify the explainability expectations of an application's end users. From a machine learning perspective, more research must be devoted to the creation of instrinsically explainable ConvNets, rather than relying solely on post-hoc explanation methods. Such a ConvNet must be aligned with the explainability requirements of its task and its users: a psoriasis diagnosis ConvNet aimed at dermatologists might



Characteristic	DenseNet121	EfficientNet-B0	InceptionResNetV2	InceptionV3	MobileNet	MobileNetV2	NASNetMobile	ResNet50	ResNet50V2	VGG16	Xception	Dermatologists
Closed comedo	0.68	0.67	0.37	0.63	0.53	0.52	0.63	0.37	0.39	0.47	0.55	0.33
Cyst	0.59	0.57	0.30	0.51	0.53	0.56	0.51	0.33	0.28	0.45	0.40	0.52
Dermatoglyph disruption	0.50	0.42	0.38	0.54	0.38	0.43	0.59	0.37	0.34	0.27	0.61	0.78
Leukotrichia	0.40	0.25	0.55	0.44	0.49	0.25	0.62	0.44	0.31	0.18	0.62	0.65
Macule	0.52	0.41	0.29	0.41	0.37	0.39	0.51	0.29	0.28	0.33	0.45	0.35
Nodule	0.65	0.60	0.28	0.53	0.45	0.51	0.54	0.35	0.36	0.44	0.47	0.55
Open comedo	0.65	0.67	0.42	0.61	0.51	0.50	0.61	0.42	0.38	0.50	0.56	0.31
Papule	0.61	0.58	0.40	0.57	0.51	0.58	0.65	0.42	0.40	0.46	0.56	0.41
Patch	0.49	0.35	0.31	0.42	0.36	0.32	0.49	0.32	0.28	0.24	0.44	0.77
Plaque	0.49	0.39	0.31	0.47	0.40	0.36	0.53	0.35	0.32	0.33	0.50	0.75
Pustule	0.72	0.70	0.52	0.67	0.60	0.65	0.70	0.44	0.48	0.58	0.62	0.34
Scale	0.56	0.44	0.36	0.51	0.46	0.41	0.55	0.42	0.36	0.36	0.53	0.63
Scar	0.58	0.59	0.29	0.55	0.44	0.51	0.58	0.31	0.30	0.42	0.52	0.40
Sun damage	0.33	0.31	0.15	0.35	0.28	0.27	0.33	0.22	0.20	0.24	0.29	0.79
Telangiectasia	0.29	0.32	0.13	0.40	0.25	0.23	0.38	0.26	0.23	0.30	0.31	0.48
Thrombosed capillaries	0.57	0.51	0.53	0.64	0.46	0.45	0.78	0.50	0.45	0.34	0.77	0.60
Mean	0.54	0.49	0.35	0.52	0.44	0.43	0.56	0.36	0.34	0.37	0.51	0.54

Figure 5: Explainability performance in terms of characteristic-level Grad-CAM sensitivity for the ConvNets (averaged across five runs) and dermatologists (averaged across eight experts). NASNetMobile and Xception outperform expert level in seven characteristics, while no ConvNet achieves expert-level performance in eight characteristics.

require high characteristic-level explainability to offer a constrative explanation against a possible differential diagnosis of atopic dermatitis, while the same ConvNet aimed at patients might require high image-level explainability to reassure the patient that all aspects of their condition are taken into consideration.

5.4 Limitations and future work

Our work has a few limitations. First, the original DermXDB dataset contains little information about the gender, age, and ethnicity of the subjects, leading to difficulties in performing an in-depth bias analysis of our benchmark. Second, the small size of the dataset limits the training capabilities of our benchmark, which may underestimate the performance of the larger ConvNets.

In future work, we plan on expanding this benchmark by using more explainability methods, such as saliency maps and LIME, to also create a benchmark of explainability methods and their performance compared to that of dermatologists. Additionally, with the increased popularity of visual transformers [Khan et al., 2022], an analysis of their Grad-CAM explainability would be of interest to the research world.

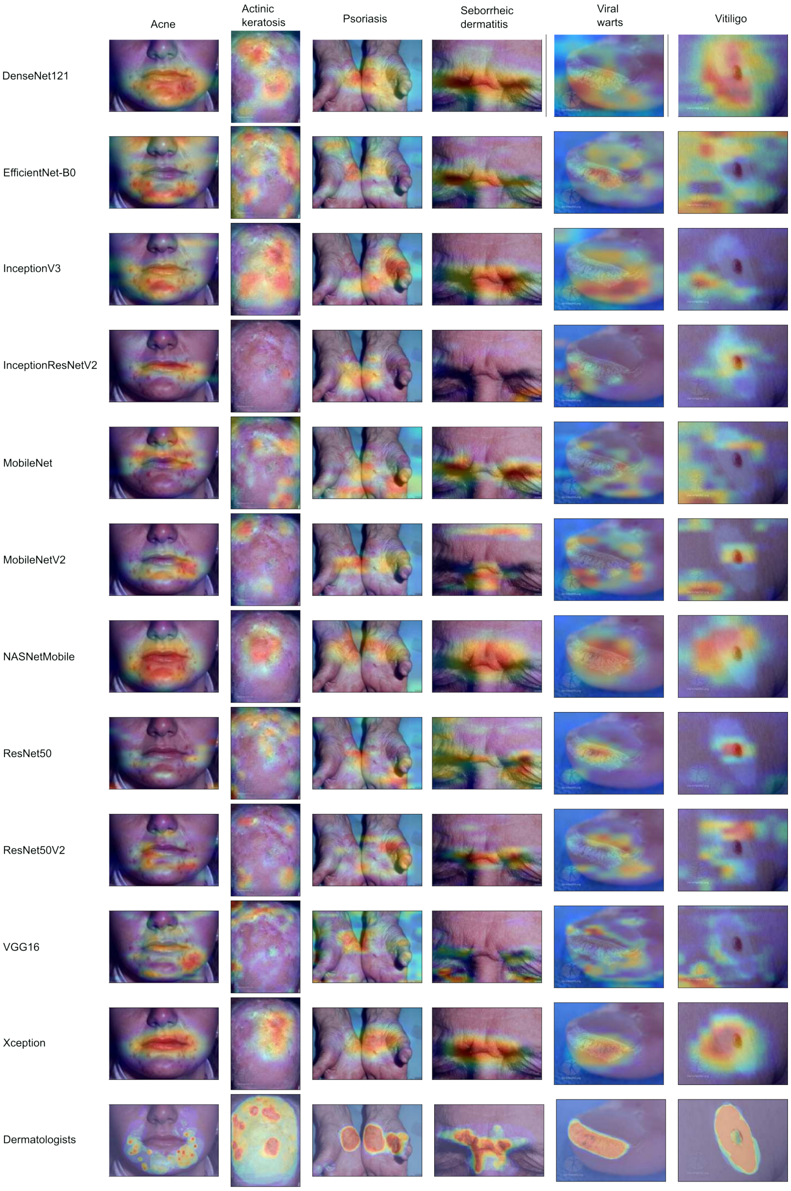


Figure 6: Example of Grad-CAM outputs for six images that were correctly diagnosed by all ConvNets. Older ConvNets, such as VGG16, ResNet50, ResNet50V2, and InceptionResNetV2, tend to focus on a single, highly indicative lesion rather than the whole affected region. More modern ConvNets, such as NASNetMobile, Xception, and EfficientNet, focus on the entire affected area. Some ConvNets overfitted during training, and focus on the watermark when diagnosing vitiligo.

6 Conclusions

In this paper, we performed a systematic literature review to identify the most used ConvNet architectures for the diagnosis of skin diseases from photographic images. We benchmarked the 11 identified ConvNets on DermXDB, a skin disease explainability dataset. Xception stands out as a highly explainable ConvNet, although NASNetMobile outperforms it on characteristic-level sensitivity. Our findings highlight the importance of explainability benchmarking, and will hopefully motivate additional studies within the field of quantitative evaluations for explainability.

Acknowledgments

Funding: RJ's work was supported in part by the Danish Innovation Fund under Grant 0153-00154A. OW's work was funded in part by the Novo Nordisk Foundation through the Center for Basic Machine Learning Research in Life Science (NNF20OC0062606). OW acknowledges support from the Pioneer Centre for AI, DNRF grant number P1.

References

World Health Organization. Working for health and growth: investing in the health workforce. 2016.

U.S. Food and Drug Administration. FDA permits marketing of artificial intelligence-based device to detect certain diabetes-related eye problems. *News Release, April*, 2018.

Alex Krizhevsky, Ilya Sutskever, and Geoffrey E Hinton. Imagenet classification with deep convolutional neural networks. *Communications of the ACM*, 60(6):84–90, 2017.

Olga Russakovsky, Jia Deng, Hao Su, Jonathan Krause, Sanjeev Satheesh, Sean Ma, Zhiheng Huang, Andrej Karpathy, Aditya Khosla, Michael Bernstein, Alexander C. Berg, and Li Fei-Fei. ImageNet Large Scale Visual Recognition Challenge. *International Journal of Computer Vision (IJCV)*, 115(3):211–252, 2015.

Jun Gao, Qian Jiang, Bo Zhou, and Daozheng Chen. Convolutional neural networks for computer-aided detection or diagnosis in medical image analysis: an overview. *Mathematical Biosciences and Engineering*, 16(6):6536–6561, 2019.

Thorsten Falk, Dominic Mai, Robert Bensch, Özgün Çiçek, Ahmed Abdulkadir, Yassine Marrakchi, Anton Böhm, Jan Deubner, Zoe Jäckel, Katharina Seiwald, et al. U-net: deep learning for cell counting, detection, and morphometry. *Nature Methods*, 16(1):67–70, 2019.

Varun Gulshan, Lily Peng, Marc Coram, Martin C Stumpe, Derek Wu, Arunachalam Narayanaswamy, Subhashini Venugopalan, Kasumi Widner, Tom Madams, Jorge Cuadros, et al. Development and validation of a deep learning algorithm for detection of diabetic retinopathy in retinal fundus photographs. *JAMA*, 316(22):2402–2410, 2016.

Richard Kijowski, Fang Liu, Francesco Caliva, and Valentina Pedoia. Deep learning for lesion detection, progression, and prediction of musculoskeletal disease. *Journal of Magnetic Resonance Imaging*, 52(6):1607–1619, 2020.

Mohammad Hesam Hesamian, Wenjing Jia, Xiangjian He, and Paul Kennedy. Deep learning techniques for medical image segmentation: achievements and challenges. *Journal of Digital Imaging*, 32:582–596, 2019.

KKD Ramesh, G Kiran Kumar, K Swapna, Debabrata Datta, and S Suman Rajest. A review of medical image segmentation algorithms. *EAI Endorsed Transactions on Pervasive Health and Technology*, 7(27):e6–e6, 2021.

Andre Esteva, Brett Kuprel, Roberto A Novoa, Justin Ko, Susan M Swetter, Helen M Blau, and Sebastian Thrun. Dermatologist-level classification of skin cancer with deep neural networks. *Nature*, 542(7639):115–118, 2017.

Jianpeng Zhang, Yutong Xie, Yong Xia, and Chunhua Shen. Attention residual learning for skin lesion classification. *IEEE Transactions on Medical Imaging*, 38(9):2092–2103, 2019.

Shunichi Jinnai, Naoya Yamazaki, Yuichiro Hirano, Yohei Sugawara, Yuichiro Ohe, and Ryuji Hamamoto. The development of a skin cancer classification system for pigmented skin lesions using deep learning. *Biomolecules*, 10 (8):1123, 2020.

Holger Andreas Haenssle, Christine Fink, Ferdinand Toberer, Julia Winkler, Wilhelm Stolz, Teresa Deinlein, Rainer Hofmann-Wellenhof, Aimilios Lallas, Steffen Emmert, Timo Buhl, et al. Man against machine reloaded: performance of a market-approved convolutional neural network in classifying a broad spectrum of skin lesions in comparison with 96 dermatologists working under less artificial conditions. *Annals of Oncology*, 31(1):137–143, 2020.

Abhijit Guha Roy, Jie Ren, Shekoofeh Azizi, Aaron Loh, Vivek Natarajan, Basil Mustafa, Nick Pawlowski, Jan Freyberg, Yuan Liu, Zach 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.

Christopher J Kelly, Alan Karthikesalingam, Mustafa Suleyman, Greg Corrado, and Dominic King. Key challenges for delivering clinical impact with artificial intelligence. *BMC Medicine*, 17(1):1–9, 2019.

Bryce Goodman and Seth Flaxman. European union regulations on algorithmic decision-making and a “right to explanation”. *AI Magazine*, 38(3):50–57, 2017.

Xiao Bai, Xiang Wang, Xianglong Liu, Qiang Liu, Jingkuan Song, Nicu Sebe, and Been Kim. Explainable deep learning for efficient and robust pattern recognition: A survey of recent developments. *Pattern Recognition*, 120: 108102, 2021.

Ramprasaath R Selvaraju, Michael Cogswell, Abhishek Das, Ramakrishna Vedantam, Devi Parikh, and Dhruv 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.

Scott M Lundberg and Su-In Lee. A unified approach to interpreting model predictions. *Advances in Neural Information Processing Systems*, 30, 2017.

Marco Tulio Ribeiro, Sameer Singh, and Carlos Guestrin. "Why should I trust you?" Explaining the predictions of any classifier. In *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pages 1135–1144, 2016.

Raluca Jalaboi, Frederik Faye, Mauricio Orbes-Arteaga, Dan Jøgensen, Ole Winther, and Alfiia Galimzianova. Dermx: an end-to-end framework for explainable automated dermatological diagnosis. *Medical Image Analysis*, page 102647, 2022.

Kenneth Thomsen, Lars Iversen, Therese Louise Titlestad, and Ole Winther. Systematic review of machine learning for diagnosis and prognosis in dermatology. *Journal of Dermatological Treatment*, 31(5):496–510, 2020. doi:https://doi.org/10.1080/09546634.2019.1682500.

Hyeon Ki Jeong, Christine Park, Ricardo Henao, and Meenal Kheterpal. Deep learning in dermatology: a systematic review of current approaches, outcomes and limitations. *JID Innovations*, page 100150, 2022.

Zhao Liu, Jiuai Sun, Lyndon Smith, Melvyn Smith, and Robert Warr. Distribution quantification on dermoscopy images for computer-assisted diagnosis of cutaneous melanomas. *Medical & Biological Engineering & Computing*, 50(5): 503–513, 2012.

Peyman Sabouri, Hamid GholamHosseini, Thomas Larsson, and John Collins. A cascade classifier for diagnosis of melanoma in clinical images. In *2014 36th Annual International Conference of the IEEE Engineering in Medicine and Biology Society*, pages 6748–6751. IEEE, 2014.

Ravneet Kaur, Peter P Albano, Justin G Cole, Jason Hagerty, Robert W LeAnder, Randy H Moss, and William V Stoecker. Real-time supervised detection of pink areas in dermoscopic images of melanoma: importance of color shades, texture and location. *Skin Research and Technology*, 21(4):466–473, 2015.

S Kefel, S Pelin Kefel, RW LeAnder, R Kaur, R Kasmi, NK Mishra, RK Rader, JG Cole, ZT Woolsey, and WV Stoecker. Adaptable texture-based segmentation by variance and intensity for automatic detection of semitranslucent and pink blush areas in basal cell carcinoma. *Skin Research and Technology*, 22(4):412–422, 2016.

Kouhei Shimizu, Hitoshi Iyatomi, M Emre Celebi, Kerri-Ann Norton, and Masaru Tanaka. Four-class classification of skin lesions with task decomposition strategy. *IEEE Transactions on Biomedical Engineering*, 62(1):274–283, 2014.

Jose Luis García Arroyo and Begoña García Zapirain. Detection of pigment network in dermoscopy images using supervised machine learning and structural analysis. *Computers in Biology and Medicine*, 44:144–157, 2014.

Vimal K Shrivastava, Narendra D Londhe, Rajendra S Sonawane, and Jasjit S Suri. Computer-aided diagnosis of psoriasis skin images with hos, texture and color features: a first comparative study of its kind. *Computer Methods and Programs in Biomedicine*, 126:98–109, 2016.

Noel CF Codella, David Gutman, M Emre Celebi, Brian Helba, Michael A Marchetti, Stephen W Dusza, Aadi Kalloo, Konstantinos Liopyris, Nabin Mishra, Harald Kittler, 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)*, pages 168–172. IEEE, 2018.

Xiaoxiao Sun, Jufeng Yang, Ming Sun, and Kai Wang. A benchmark for automatic visual classification of clinical skin disease images. In *European Conference on Computer Vision*, pages 206–222. Springer, 2016.

DermNetNZ. DermNetNZ. `https://dermnetnz.org/`, 2021. Accessed: 2021-04-01.

Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. Imagenet: A large-scale hierarchical image database. In *2009 IEEE Conference on Computer Vision and Pattern Recognition*, pages 248–255. Ieee, 2009.

P Tschandl, H Kittler, and G Argenziano. A pretrained neural network shows similar diagnostic accuracy to medical students in categorizing dermatoscopic images after comparable training conditions. *British Journal of Dermatology*, 177(3):867–869, 2017.

Seung Seog Han, Gyeong Hun Park, Woohyung Lim, Myoung Shin Kim, Jung Im Na, Ilwoo Park, and Sung Eun 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.

G Reshma, Chiai Al-Atroshi, V Kumar Nassa, B Geetha, Gurram Sunitha, Mohammad Gouse Galety, and S Neelakandan. Deep learning-based skin lesion diagnosis model using dermoscopic images. *Intelligent Automation and Soft Computing*, 31(1):621–634, 2022.

Yading Yuan, Ming Chao, and Yeh-Chi Lo. Automatic skin lesion segmentation using deep fully convolutional networks with jaccard distance. *IEEE Transactions on Medical Imaging*, 36(9):1876–1886, 2017.

Huisi Wu, Shihuai Chen, Guilian Chen, Wei Wang, Baiying Lei, and Zhenkun Wen. Fat-net: Feature adaptive transformers for automated skin lesion segmentation. *Medical Image Analysis*, 76:102327, 2022.

Ramsha Baig, Maryam Bibi, Anmol Hamid, Sumaira Kausar, and Shahzad Khalid. Deep learning approaches towards skin lesion segmentation and classification from dermoscopic images-a review. *Current Medical Imaging*, 16(5): 513–533, 2020.

Philipp Tschandl, Christoph Rinner, Zoe Apalla, Giuseppe Argenziano, Noel Codella, Allan Halpern, Monika Janda, Aimilios Lallas, Caterina Longo, Josep Malvehy, et al. Human–computer collaboration for skin cancer recognition. *Nature Medicine*, 26(8):1229–1234, 2020.

Masaya Tanaka, Atsushi Saito, Kosuke Shido, Yasuhiro Fujisawa, Kenshi Yamasaki, Manabu Fujimoto, Kohei Murao, Youichirou Ninomiya, Shin'ichi Satoh, and Akinobu Shimizu. Classification of large-scale image database of various skin diseases using deep learning. *International Journal of Computer Assisted Radiology and Surgery*, 16:1875–1887, 2021.

Ribana Roscher, Bastian Bohn, Marco F Duarte, and Jochen Garcke. Explainable machine learning for scientific insights and discoveries. *IEEE Access*, 8:42200–42216, 2020.

Amitojdeep Singh, Sourya Sengupta, and Vasudevan Lakshminarayanan. Explainable deep learning models in medical image analysis. *Journal of Imaging*, 6(6):52, 2020.

Stan Lipovetsky and Michael Conklin. Analysis of regression in game theory approach. *Applied Stochastic Models in Business and Industry*, 17(4):319–330, 2001.

Mara Graziani, Iam Palatnik de Sousa, Marley MBR Vellasco, Eduardo Costa da Silva, Henning Müller, and Vincent Andrearczyk. Sharpening local interpretable model-agnostic explanations for histopathology: Improved understandability and reliability. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, pages 540–549. Springer, 2021.

Karen Simonyan and Andrew Zisserman. Very deep convolutional networks for large-scale image recognition. In *International Conference on Learning Representations*, 2015.

Jost Tobias Springenberg, Alexey Dosovitskiy, Thomas Brox, and Martin Riedmiller. Striving for simplicity: The all convolutional net. *arXiv preprint arXiv:1412.6806*, 2014.

Sebastian Bach, Alexander Binder, Grégoire Montavon, Frederick Klauschen, Klaus-Robert Müller, and Wojciech Samek. On pixel-wise explanations for non-linear classifier decisions by layer-wise relevance propagation. *PloS One*, 10(7):e0130140, 2015.

Sérgio Pereira, Raphael Meier, Victor Alves, Mauricio Reyes, and Carlos A Silva. Automatic brain tumor grading from mri data using convolutional neural networks and quality assessment. In *Understanding and Interpreting Machine Learning in Medical Image Computing Applications*, pages 106–114. Springer, 2018.

Kyle Young, Gareth Booth, Becks Simpson, Reuben Dutton, and Sally Shrapnel. Deep neural network or dermatologist? In *Interpretability of Machine Intelligence in Medical Image Computing and Multimodal Learning for Clinical Decision Support*, pages 48–55. Springer, 2019.

Tobias Hepp, Dominik Blum, Karim Armanious, Bernhard Schoelkopf, Darko Stern, Bin Yang, and Sergios Gatidis. Uncertainty estimation and explainability in deep learning-based age estimation of the human brain: Results from the german national cohort mri study. *Computerized Medical Imaging and Graphics*, 92:101967, 2021.

Raluca Jalaboi, Ole Winther, and Alfiia Galimzianova. Explainable image quality assessments in teledermatological photography. *Telemedicine and e-Health*, 2023.

William R Crum, Oscar Camara, and Derek L G Hill. Generalized overlap measures for evaluation and validation in medical image analysis. *IEEE Transactions on Medical Imaging*, 25(11):1451–1461, 2006.

François Chollet. keras. `https://github.com/fchollet/keras`, 2015.

Christian Szegedy, Wei Liu, Yangqing Jia, Pierre Sermanet, Scott Reed, Dragomir Anguelov, Dumitru Erhan, Vincent Vanhoucke, and Andrew Rabinovich. Going deeper with convolutions. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 1–9, 2015.

Artem Babenko and Victor Lempitsky. Aggregating deep convolutional features for image retrieval. *arXiv preprint arXiv:1510.07493*, 2015.

Andrew Howard, Mark Sandler, Grace Chu, Liang-Chieh Chen, Bo Chen, Mingxing Tan, Weijun Wang, Yukun Zhu, Ruoming Pang, Vijay Vasudevan, et al. Searching for mobilenetv3. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 1314–1324, 2019.

Jie Hu, Li Shen, and Gang Sun. Squeeze-and-excitation networks. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 7132–7141, 2018.

Xiangyu Zhang, Xinyu Zhou, Mengxiao Lin, and Jian Sun. Shufflenet: An extremely efficient convolutional neural network for mobile devices. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 6848–6856, 2018.

Gao Huang, Zhuang Liu, Laurens Van Der Maaten, and Kilian Q Weinberger. Densely connected convolutional networks. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 4700–4708, 2017.

Mingxing Tan and Quoc Le. Efficientnet: Rethinking model scaling for convolutional neural networks. In *International Conference on Machine Learning*, pages 6105–6114. PMLR, 2019.

Christian Szegedy, Sergey Ioffe, Vincent Vanhoucke, and Alexander A Alemi. Inception-v4, inception-resnet and the impact of residual connections on learning. In *Thirty-first AAAI Conference on Artificial Intelligence*, 2017.

Christian Szegedy, Vincent Vanhoucke, Sergey Ioffe, Jon Shlens, and Zbigniew Wojna. Rethinking the inception architecture for computer vision. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 2818–2826, 2016.

Andrew G Howard, Menglong Zhu, Bo Chen, Dmitry Kalenichenko, Weijun Wang, Tobias Weyand, Marco Andreetto, and Hartwig Adam. Mobilenets: Efficient convolutional neural networks for mobile vision applications. *arXiv preprint arXiv:1704.04861*, 2017.

Mark Sandler, Andrew Howard, Menglong Zhu, Andrey Zhmoginov, and Liang-Chieh Chen. Mobilenetv2: Inverted residuals and linear bottlenecks. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 4510–4520, 2018.

Barret Zoph, Vijay Vasudevan, Jonathon Shlens, and Quoc V Le. Learning transferable architectures for scalable image recognition. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 8697–8710, 2018.

Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 770–778, 2016a.

Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Identity mappings in deep residual networks. In *European Conference on Computer Vision*, pages 630–645. Springer, 2016b.

François Chollet. Xception: Deep learning with depthwise separable convolutions. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 1251–1258, 2017.

1st Lt Pushkar Aggarwal. Data augmentation in dermatology image recognition using machine learning. *Skin Research and Technology*, 25(6):815–820, 2019.

Philippe M Burlina, Neil J Joshi, Elise Ng, Seth D Billings, Alison W Rebman, and John N Aucott. Automated detection of erythema migrans and other confounding skin lesions via deep learning. *Computers in Biology and Medicine*, 105:151–156, 2019.

Xin-yu Zhao, Xian Wu, Fang-fang Li, Yi Li, Wei-hong Huang, Kai Huang, Xiao-yu He, Wei Fan, Zhe Wu, Ming-liang Chen, et al. The application of deep learning in the risk grading of skin tumors for patients using clinical images. *Journal of Medical Systems*, 43(8):1–7, 2019.

Philippe M Burlina, Neil J Joshi, Phil A Mathew, William Paul, Alison W Rebman, and John N Aucott. Ai-based detection of erythema migrans and disambiguation against other skin lesions. *Computers in Biology and Medicine*, 125:103977, 2020.

YPH Chin, ZY Hou, MY Lee, HM Chu, HH Wang, YT Lin, A Gittin, SC Chien, PA Nguyen, LC Li, et al. A patient-oriented, general-practitioner-level, deep-learning-based cutaneous pigmented lesion risk classifier on a smartphone. *The British Journal of Dermatology*, 182(6):1498–1500, 2020.

Seung Seog Han, Ilwoo Park, Sung Eun Chang, Woohyung Lim, Myoung Shin Kim, Gyeong Hun Park, Je Byeong Chae, Chang Hun Huh, and Jung-Im Na. Augmented intelligence dermatology: deep neural networks empower medical professionals in diagnosing skin cancer and predicting treatment options for 134 skin disorders. *Journal of Investigative Dermatology*, 140(9):1753–1761, 2020.

Yuan Liu, Ayush Jain, Clara Eng, David H Way, Kang Lee, Peggy Bui, Kimberly Kanada, Guilherme de Oliveira Marinho, Jessica Gallegos, Sara Gabriele, et al. A deep learning system for differential diagnosis of skin diseases. *Nature Medicine*, 26(6):900–908, 2020.

Shuang Zhao, Bin Xie, Yi Li, X-y Zhao, Yehong Kuang, Juan Su, X-y He, Xian Wu, Wei Fan, Kai Huang, et al. Smart identification of psoriasis by images using convolutional neural networks: a case study in china. *Journal of the European Academy of Dermatology and Venereology*, 34(3):518–524, 2020.

Haijing Wu, Heng Yin, Haipeng Chen, Moyuan Sun, Xiaoqing Liu, Yizhou Yu, Yang Tang, Hai Long, Bo Zhang, Jing Zhang, et al. A deep learning-based smartphone platform for cutaneous lupus erythematosus classification assistance: Simplifying the diagnosis of complicated diseases. *Journal of the American Academy of Dermatology*, 85 (3):792–793, 2021.

Pushkar Aggarwal and Francis A Papay. Artificial intelligence image recognition of melanoma and basal cell carcinoma in racially diverse populations. *Journal of Dermatological Treatment*, 33(4):2257–2262, 2022.

Wei Ba, Huan Wu, Wei W Chen, Shu H Wang, Zi Y Zhang, Xuan J Wei, Wen J Wang, Lei Yang, Dong M Zhou, Yi X Zhuang, et al. Convolutional neural network assistance significantly improves dermatologists' diagnosis of cutaneous tumours using clinical images. *European Journal of Cancer*, 169:156–165, 2022.

Sk Imran Hossain, Jocelyn de Goër de Herve, Md Shahriar Hassan, Delphine Martineau, Evelina Petrosyan, Violaine Corbin, Jean Beytout, Isabelle Lebert, Jonas Durand, Irene Carravieri, et al. Exploring convolutional neural networks with transfer learning for diagnosing lyme disease from skin lesion images. *Computer Methods and Programs in Biomedicine*, 215:106624, 2022.

Jens Hüsers, Guido Hafer, Jan Heggemann, Stefan Wiemeyer, Mareike Przysucha, Joachim Dissemond, Maurice Moelleken, Cornelia Erfurt-Berge, and Ursula Hübner. Automatic classification of diabetic foot ulcer images–a transfer-learning approach to detect wound maceration. In *Informatics and Technology in Clinical Care and Public Health*, pages 301–304. IOS Press, 2022.

Tom J Liu, Mesakh Christian, Yuan-Chia Chu, Yu-Chun Chen, Che-Wei Chang, Feipei Lai, and Hao-Chih Tai. A pressure ulcers assessment system for diagnosis and decision making using convolutional neural networks. *Journal of the Formosan Medical Association*, 121(11):2227–2236, 2022.

Leila Malihi, Jens Hüsers, Mats L Richter, Maurice Moelleken, Mareike Przysucha, Dorothee Busch, Jan Heggemann, Guido Hafer, Stefan Wiemeyer, Gunther Heidemann, et al. Automatic wound type classification with convolutional neural networks. *Advances in Informatics, Management and Technology in Healthcare*, 295:281, 2022.

A Munthuli, J Intanai, P Tossanuch, P Pooprasert, P Ingpochai, S Boonyasatian, K Kittithammo, P Thammarach, T Boonmak, S Khaengthanyakan, et al. Extravasation screening and severity prediction from skin lesion image using deep neural networks. In *2022 44th Annual International Conference of the IEEE Engineering in Medicine & Biology Society (EMBC)*, pages 1827–1833. IEEE, 2022.

Ruiyan Ni, Ta Zhou, Ge Ren, Yuanpeng Zhang, Dongrong Yang, Victor CW Tam, Wan Shun Leung, Hong Ge, Shara WY Lee, and Jing Cai. Deep learning-based automatic assessment of radiation dermatitis in patients with nasopharyngeal carcinoma. *International Journal of Radiation Oncology* Biology* Physics*, 113(3):685–694, 2022.

Veysel Harun Sahin, Ismail Oztel, and Gozde Yolcu Oztel. Human monkeypox classification from skin lesion images with deep pre-trained network using mobile application. *Journal of Medical Systems*, 46(11):79, 2022.

Meng Xia, Meenal K Kheterpal, Samantha C Wong, Christine Park, William Ratliff, Lawrence Carin, and Ricardo Henao. Lesion identification and malignancy prediction from clinical dermatological images. *Scientific Reports*, 12 (1):15836, 2022.

Jiancun Zhou, Zheng Wu, Zixi Jiang, Kai Huang, Kehua Guo, and Shuang Zhao. Background selection schema on deep learning-based classification of dermatological disease. *Computers in Biology and Medicine*, 149:105966, 2022.

Hang Zhang and Tianyi Ma. Acne detection by ensemble neural networks. *Sensors*, 22(18):6828, 2022.

Alexander Nast, Chris EM Griffiths, Roderick Hay, Wolfram Sterry, and Jean L Bolognia. The 2016 international league of dermatological societies' revised glossary for the description of cutaneous lesions. *British Journal of Dermatology*, 174(6):1351–1358, 2016.

Salman Khan, Muzammal Naseer, Munawar Hayat, Syed Waqas Zamir, Fahad Shahbaz Khan, and Mubarak Shah. Transformers in vision: A survey. *ACM Computing Surveys (CSUR)*, 54(10s):1–41, 2022.

Appendix

Table A1 presents statistics for the proprietary clinical dataset used in the hyper-parameter search and the pre-training step. Table A2 shows the best performing list of parameters identified for each ConvNet. The search space consisted of the following values for each hyperparameter:

- **Rotation:** 10, 20
- **Shear:** 0.00, 0.25, 0.50
- **Zoom:** 0.25, 0.5
- **Brightness ranges:** 0.00-0.50, 0.00-0.25, 0.50-1.00, 0.50-1.50, 0.75-1.25
- **Learning rate:** 0.01, 0.001, 0.0001
- **Last fixed layer:** last convolutional layer, second to last convolutional block
- **Epochs:** 10, 25, 50, 75

Table A1: Dataset statistics for the proprietary pre-training clinical dataset.

Diagnosis	Training	Validation
Acne	832	245
Actinic keratosis	132	33
Psoriasis	771	204
Seborrheic dermatitis	88	25
Viral warts	509	97
Vitiligo	141	37

Table A2: Optimal list of hyperparameters for each ConvNet, as identified after a hyper-parameter search.

ConvNet	Rotation	Shear	Zoom	Brightness	Learning rate	Last fixed layer	Epochs
DenseNet121	20	0.50	0.50	[0.50, 1.50]	0.0001	conv5_block14_concat	75
EfficientNet-B0	20	0.25	0.50	[0.50, 1.50]	0.0001	block6d_add	50
InceptionV3	20	0.50	0.50	[0.50, 1.50]	0.001	activation288	50
InceptionResNetv2	20	0.25	0.50	[0.75, 1.25]	0.0001	block8_9_ac	50
MobileNet	10	0.50	0.50	[0.50, 1.50]	0.0001	conv_pw_12_relu	50
MobileNetV2	10	0.25	0.50	[0.50, 1.50]	0.0001	block_15_add	75
NASNetMobile	20	0.25	0.50	[0.50, 1.00]	0.0001	normal_concat_11	75
ResNet50	20	0.50	0.50	[0.50, 1.50]	0.0001	conv5_block3_out	50
ResNet50V2	20	0.25	0.25	[0.50, 1.00]	0.001	post_relu	75
VGG16	10	0.00	0.25	[0.50, 1.00]	0.01	block5_pool	75
Xception	10	0.25	0.50	[0.50, 1.50]	0.001	block14_sepconv2_act	50

Table A3: Diagnostic performance of ConvNets in terms macro F1-score, sensitivity, and specificity on the validation subset of the proprietary clinical dataset (average ± standard deviation across five runs).

ConvNet	F1-score	Sensitivity	Specificity
DenseNet121	0.80 ± 0.01	0.79 ± 0.01	0.98 ± 0.00
EfficientNet-B0	0.77 ± 0.01	0.78 ± 0.01	0.97 ± 0.00
InceptionV3	0.76 ± 0.02	0.74 ± 0.02	0.96 ± 0.00
InceptionResNetV2	0.73 ± 0.02	0.73 ± 0.02	0.97 ± 0.00
MobileNet	0.72 ± 0.02	0.71 ± 0.02	0.96 ± 0.00
MobileNetV2	0.72 ± 0.03	0.73 ± 0.02	0.96 ± 0.00
NASNetMobile	0.67 ± 0.04	0.64 ± 0.02	0.95 ± 0.01
ResNet50	0.70 ± 0.01	0.68 ± 0.02	0.96 ± 0.00
ResNet50V2	0.76 ± 0.01	0.75 ± 0.02	0.96 ± 0.00
VGG16	0.66 ± 0.03	0.67 ± 0.01	0.95 ± 0.00
Xception	0.82 ± 0.03	0.82 ± 0.02	0.97 ± 0.00

Table A4: Diagnostic performance of ConvNets (average ± standard deviation across five runs) and dermatologists (average ± standard deviation across eight dermatologists) in terms of sensitivity on the DermXDB holdout set, split by diagnosis. Several ConvNets achieve expert-level sensitivity on multiple classes (in **bold**).

	Acne	**Actinic keratosis**	**Psoriasis**	**Seborrheic dermatitis**	**Viral warts**	**Vitiligo**
ConvNets						
DenseNet121	0.85 ± 0.03	**0.52 ± 0.10**	0.71 ± 0.03	**0.76 ± 0.05**	**0.91 ± 0.03**	0.66 ± 0.03
EfficientNet-B0	0.71 ± 0.08	**0.43 ± 0.09**	0.63 ± 0.07	**0.64 ± 0.13**	0.66 ± 0.05	**0.84 ± 0.13**
InceptionV3	0.83 ± 0.06	**0.55 ± 0.15**	0.62 ± 0.08	**0.54 ± 0.13**	**0.70 ± 0.10**	0.73 ± 0.12
InceptionResNetV2	0.70 ± 0.06	**0.41 ± 0.12**	0.62 ± 0.06	**0.67 ± 0.11**	0.43 ± 0.12	0.74 ± 0.07
MobileNet	0.76 ± 0.07	**0.48 ± 0.22**	0.60 ± 0.26	**0.66 ± 0.08**	0.47 ± 0.17	0.68 ± 0.07
MobileNetV2	**0.88 ± 0.08**	0.14 ± 0.06	0.21 ± 0.10	**0.63 ± 0.22**	0.33 ± 0.14	0.61 ± 0.23
NASNetMobile	0.79 ± 0.07	0.24 ± 0.11	0.33 ± 0.09	**0.48 ± 0.09**	0.42 ± 0.03	0.48 ± 0.13
ResNet50	0.79 ± 0.05	**0.40 ± 0.16**	0.69 ± 0.10	**0.74 ± 0.07**	0.54 ± 0.13	0.80 ± 0.04
ResNet50V2	**0.85 ± 0.10**	**0.55 ± 0.13**	0.58 ± 0.04	**0.61 ± 0.08**	**0.74 ± 0.07**	0.71 ± 0.02
VGG16	0.74 ± 0.10	**0.62 ± 0.09**	0.65 ± 0.06	**0.44 ± 0.18**	0.54 ± 0.11	0.76 ± 0.07
Xception	**0.89 ± 0.05**	**0.52 ± 0.08**	0.72 ± 0.06	**0.61 ± 0.11**	**0.81 ± 0.05**	0.82 ± 0.04
Dermatologists						
Average	0.95 ± 0.03	0.67 ± 0.18	0.88 ± 0.06	0.59 ± 0.11	0.88 ± 0.09	0.92 ± 0.05

Table A5: Diagnostic performance of ConvNets (average ± standard deviation across five runs) and dermatologists (average ± standard deviation across eight dermatologists) in terms of specificity on the DermXDB holdout set, split by diagnosis. Several ConvNets achieve expert-level specificity (in **bold**).

	Acne	Actinic keratosis	Psoriasis	Seborrheic dermatitis	Viral warts	Vitiligo
ConvNets						
DenseNet121	0.93 ± 0.01	0.97 ± 0.01	0.92 ± 0.01	**0.91 ± 0.02**	0.97 ± 0.01	0.98 ± 0.01
EfficientNet-B0	**0.93 ± 0.06**	0.96 ± 0.02	0.91 ± 0.01	**0.89 ± 0.03**	0.95 ± 0.01	0.95 ± 0.01
InceptionV3	0.92 ± 0.03	0.92 ± 0.02	**0.92 ± 0.03**	**0.91 ± 0.06**	0.96 ± 0.02	**0.97 ± 0.03**
InceptionResNetV2	0.94 ± 0.02	0.97 ± 0.01	0.86 ± 0.03	**0.86 ± 0.02**	0.97 ± 0.01	0.92 ± 0.04
MobileNet	0.91 ± 0.05	0.96 ± 0.02	**0.88 ± 0.07**	**0.87 ± 0.08**	0.97 ± 0.02	0.94 ± 0.02
MobileNetV2	0.66 ± 0.15	**0.99 ± 0.01**	0.98 ± 0.03	**0.79 ± 0.11**	**1.00 ± 0.00**	**0.94 ± 0.07**
NASNetMobile	0.65 ± 0.04	0.97 ± 0.01	0.96 ± 0.01	**0.86 ± 0.02**	0.96 ± 0.03	0.96 ± 0.03
ResNet50	0.93 ± 0.02	0.99 ± 0.01	0.89 ± 0.04	**0.86 ± 0.04**	0.97 ± 0.02	0.96 ± 0.01
ResNet50V2	0.90 ± 0.06	0.95 ± 0.03	0.92 ± 0.01	**0.90 ± 0.04**	0.96 ± 0.02	0.97 ± 0.01
VGG16	0.90 ± 0.07	0.92 ± 0.03	**0.90 ± 0.05**	**0.95 ± 0.04**	0.97 ± 0.02	0.92 ± 0.03
Xception	0.91 ± 0.04	0.98 ± 0.01	**0.93 ± 0.02**	**0.92 ± 0.02**	0.97 ± 0.02	0.96 ± 0.03
Dermatologists						
Average	0.99 ± 0.01	1.00 ± 0.00	0.96 ± 0.02	0.99 ± 0.01	1.00 ± 0.00	1.00 ± 0.00

APPENDIX C

Explainable Image Quality Assessments in Teledermatological Photography

Published in the Telemedicine and e-Health Journal, February 2023.

Open camera or QR reader and
scan code to access this article
and other resources online.

Explainable Image Quality Assessments in Teledermatological Photography

*Raluca Jalaboi, MS,[1,2] Ole Winther, Prof.[1,3–5]
and Alfiia Galimzianova, PhD[2]*

*[1]Department of Applied Mathematics and Computer Science, Technical University of Denmark, Lyngby, Denmark.
[2]Medable A/S, Copenhagen, Denmark.
[3]Pioneer Centre for AI, Copenhagen, Denmark.
[4]Bioinformatics Centre, Department of Biology, University of Copenhagen, Copenhagen, Denmark.
[5]Center for Genomic Medicine, Rigshospitalet, Copenhagen University Hospital, Copenhagen, Denmark.*

Abstract

Background and Objectives: *Image quality is a crucial factor in the effectiveness and efficiency of teledermatological consultations. However, up to 50% of images sent by patients have quality issues, thus increasing the time to diagnosis and treatment. An automated, easily deployable, explainable method for assessing image quality is necessary to improve the current teledermatological consultation flow. We introduce ImageQX, a convolutional neural network for image quality assessment with a learning mechanism for identifying the most common poor image quality explanations: bad framing, bad lighting, blur, low resolution, and distance issues.*

Methods: *ImageQX was trained on 26,635 photographs and validated on 9,874 photographs, each annotated with image quality labels and poor image quality explanations by up to 12 board-certified dermatologists. The photographic images were taken between 2017 and 2019 using a mobile skin disease tracking application accessible worldwide.*

Results: *Our method achieves expert-level performance for both image quality assessment and poor image quality explanation. For image quality assessment, ImageQX obtains a macro F1-score of 0.73 ± 0.01, which places it within standard deviation of the pairwise inter-rater F1-score of 0.77 ± 0.07. For poor image quality explanations, our method obtains F1-scores of between 0.37 ± 0.01 and 0.70 ± 0.01, similar to the inter-rater pairwise F1-score of between 0.24 ± 0.15 and 0.83 ± 0.06. Moreover, with a size of only 15 MB, ImageQX is easily deployable on mobile devices.*

Conclusion: *With an image quality detection performance similar to that of dermatologists, incorporating ImageQX into the teledermatology flow can enable a better, faster flow for remote consultations.*

Keywords: *teledermatology, image quality, artificial intelligence, deep learning, explainability, telemedicine*

Introduction

Within the past 2 years, consumers facing teledermatological consultations have become much more common owing to the SARS CoV-2 (COVID-19) pandemic and associated worldwide isolation measures.[1] Teledermatological consultations are carried out increasingly more often via teledermatology mobile applications that require patients to photograph their skin lesions using their mobile devices, such as smartphones and tablets, and send them to dermatologists who will then diagnose the depicted skin condition remotely.[2,3] To achieve similar quality of care to an in-person consultation, high-quality images are paramount.[2,3] However, this is rarely the

DOI: 10.1089/tmj.2022.0405

case: up to 50% of patients send images taken under poor lighting conditions, that are not centered on the lesion, or that are blurry.[4,5]

When dealing with low-quality images, two main approaches exist: image denoising and image quality detection. Image denoising processes and reconstructs noisy images such that the noise is either reduced or entirely removed. Many denoising methods introduce new artifacts into the images or obfuscate characteristics critical for diagnosis.[6] Therefore, in this article we focus on image quality detection. By detecting low-quality images directly on the patient's mobile device, we can instruct them to retake the picture in a way that improves the quality to an acceptable level. We can thus reduce the evaluation burden on dermatologists while at the same time reducing the time to diagnosis and treatment.

Several methods for image quality detection have been proposed in the literature. Kim and Lee introduce DeepIQ,[7] a deep neural network that can identify noisy sections in an image, and compare the resulting noise maps with human assessments. Bianco et al propose DeepBIQ,[8] a convolutional neural network for identifying low-quality images, and report near human-level results on smartphone photos from the LIVE In the Wild challenge dataset.[9] Madhusudana et al develop CONTRIQUE,[10] a contrastive deep learning system for creating generalizable representations using unlabeled image quality datasets. One common issue for all methods is the lack of a reference standard label, which limits both their training and validation rigor. Because of this reason, they often use unsupervised training methods and limit validation to qualitative assessment.

Within teledermatology, Vodrahalli et al proposed a classical machine learning image quality classifier.[5] Their method provides patients with explanations for the quality assessments through automated classical computer vision methods for detecting blur, lighting, and zoom issues in an image. However, this method has several limitations: it cannot handle cases where only the background is blurry or with poor lighting, it cannot detect lesion framing issues, and it cannot discard images containing no skin.

The lack of explainability is regarded as one of the biggest obstacles toward the adoption of automated methods in medical practice.[11–13] Gradient-based class activation maps (Grad-CAM)[14] is the most common explainability method in medical computer vision owing to its ease of use, intuitive output, and low computational requirements. Grad-CAM creates CAMs on a given convolutional layer using the backpropagation gradients—the higher the gradient, the more important the region is to the final classification.

In this study, we introduce ImageQX, a convolutional neural network-based method for detecting image quality. Our novel approach uses image quality evaluations obtained from dermatologists in a teledermatology setting to learn the image quality required for a successful remote consultation. *Figure 1*illustrates the ImageQX architecture, which learns the image quality and its explanations in an end-to-end manner. ImageQX was trained and validated on 36,509 images collected using a skin lesion progression tracking mobile application. Images were labeled by up to 12 board-certified dermatologists. We evaluate the network performance with regard to the reference standard, and we obtain a macro F1-score of 0.73 for image quality assessment, with the per-explanation performance between 0.37 and 0.71. Moreover, ImageQX occupies only 15 MB, making it ideal for deploying on mobile devices as a prefiltering step during data collection.

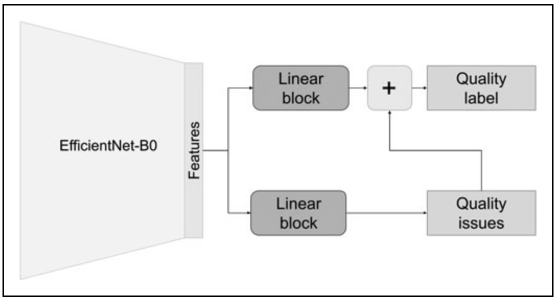


Fig. 1. ImageQX network architecture. To facilitate deployment on mobile devices, we use the lightweight EfficientNet-Bo architecture as a feature extractor. A linear block, composed of a linear layer, batch normalization, and a dropout layer, is used to parse these features before predicting poor image quality explanations, that is, *bad framing*, *bad light*, *blurry*, *low resolution*, and *too far away*. Another similar linear block parses the image features and then concatenates them with the poor image quality explanations to predict the image quality label.

Methods

A total of 36,509 images were collected between 2017 and 2019, using Imagine,[15] a skin disease tracking mobile application available worldwide. Self-reported user ages range between 18 and 80 years, and self-reported sex showing a distribution of 49% men, 47% women, and 4% other. Users span 146 countries, with images from Ukraine, United Kingdom, United States, Georgia, Russia, Albania, Kazakhstan, India, Denmark, South Africa, Bulgaria, and Israel making up 45% of the dataset. Images cover a wide variety of body parts. Self-reported body part tags show that faces, arms, elbows, legs, and groin comprise the majority of images. All data was anonymized a priori and did not involve human subjects. 45 CFR part 46 does not apply, and thus an independent ethics committee approval was not applicable for this research.

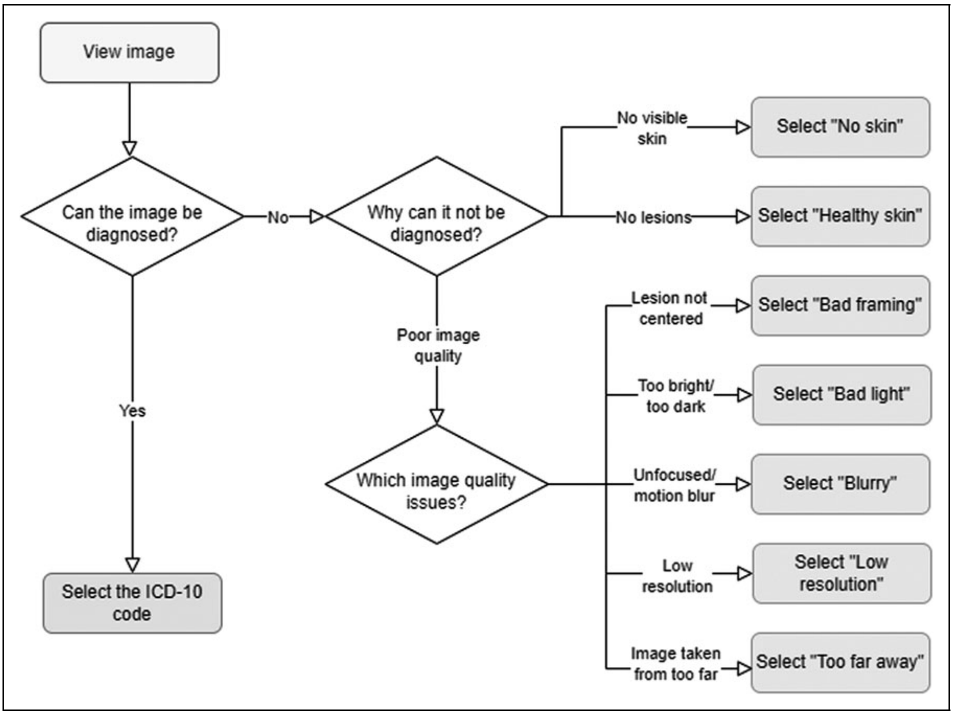


Fig. 2. Labeling protocol for the ImageQX training and validation dataset. Dermatologists start by assessing whether or not the image can be diagnosed. If the image can be assessed, they diagnose it using an ICD-10 code. Otherwise, if there is no visible skin or if there are no visible lesions in the picture, the dermatologists discard the image as *no skin* or *healthy skin*, respectively. Finally, if the image cannot be evaluated because of poor quality, they select one of the five investigated poor image quality explanations.

Each image was evaluated by up to 12 board-certified dermatologists using an in-house labeling tool. Dermatologists diagnosed each image with an *International Classification of Diseases, 10th Revision* (ICD-10) code[16] whenever a lesion was present in the image and was depicted with a sufficient quality, or alternatively with one of three nonlesion labels: *poor quality* when the image quality detracted from their ability to diagnose, *healthy skin* whenever no lesions were visible, or *no skin* for images that had no dermatological relevance. *Poor quality* images were additionally tagged with

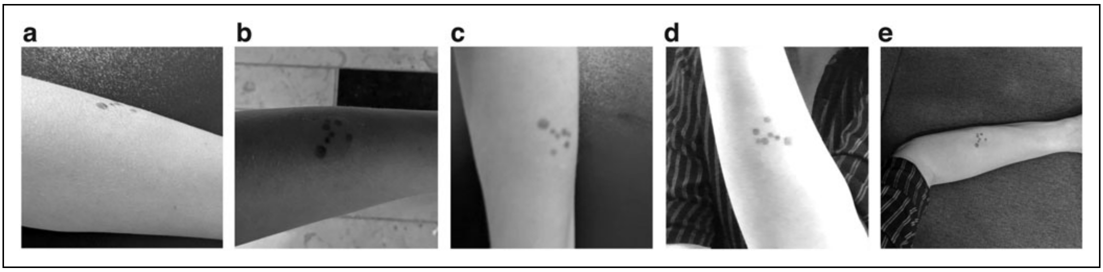


Fig. 3. Illustration of poor image quality explanations that can be detected by ImageQX. **(a)** *Bad framing*: the image was not centered on the lesion. **(b)** *Bad light*: the lighting conditions in which the image was taken were too dark or too bright. **(c)** *Blurry*: the image is not focused on the lesion, masking out its details. **(d)** *Low resolution*: the image was taken with a low-resolution camera and few details can be discerned. **(e)** *Too far away*: few lesion details could be seen owing to the distance from the camera. Images courtesy of the authors.

poor quality explanations: *bad framing* for images not centered on the lesion, *bad light* for images that are too bright or too dark, *blurry* for images suffering from motion blur or inadequate focus, *low resolution* for images taken with a low-resolution camera, or *too far away* for images where the picture was taken from afar and no details could be discerned. *Figure 2* outlines the protocol dermatologists followed when labeling the data, whereas *Figure 3* illustrates each poor image quality explanation included in the dataset.

We evaluate the performance of the raters and the network using sensitivity:

$$Se = \frac{TP}{TP + FN},$$

specificity:

$$Sp = \frac{TN}{TN + FP},$$

and F1-score:

$$F1 = \frac{2TP}{2TP + FP + FN},$$

where TP, FP, and FN denote the true positives, false positives, and false negatives, respectively. The inter-rater pairwise F1-score is calculated as the average of all dermatologist pairs, where one dermatologist is considered the reference standard whereas the other is considered the prediction. For evaluating the network performance, we calculate the macro F1-score, that is, we average the F1-scores for each class.

During training, we parsed the dermatologist evaluations into four classes by merging all ICD-10 evaluations into the *lesion* class. We used plurality label fusion, that is, the class selected by most dermatologists, for defining the image quality class for each image. Alongside assessing whether the image can be evaluated, our proposed method also offers explanations to the poor quality images. To obtain the reference standard for poor image quality explanations, we chose to mark explanations as relevant if at least one dermatologist discarded an image with that explanation. *Table 1* provides the distribution of labels within the dataset, whereas *Table 2* details the distribution of poor image quality explanations. Higher agreement is achieved on *lesion* and *no skin*, whereas low agreement between raters can be seen for *healthy skin* and *poor quality*. Poor image quality explanations display low inter-rater agreements, with *blurry* being the only one achieving an inter-rater pairwise F1-score of >0.80.

Table 1. Distribution of Image Quality Labels Over the Training and Test Sets, Including the Pairwise Inter-Rater Agreement Calculated as the Pairwise F1-Score

CLASS	TRAIN IMAGE COUNT	TEST IMAGE COUNT	PAIRWISE TRAIN F1	PAIRWISE TEST F1
Lesion	17,534	4,803	0.86 ± 0.03	0.84 ± 0.08
No skin	461	265	0.93 ± 0.03	0.92 ± 0.04
Healthy skin	3,903	2,421	0.62 ± 0.10	0.65 ± 0.10
Poor quality	4,737	2,385	0.63 ± 0.08	0.67 ± 0.07
Mean	6658.75	2468.5	0.76 ± 0.06	0.77 ± 0.07

The ImageQX architecture is inspired by the DermX architecture introduced by Jalaboi et al to intrinsically learn the expert explanations, as illustrated in *Figure 1*.[17] EfficientNet-B0 was used as the feature extractor to increase the image processing speed and reduce the network size.[18] To increase the convergence speed, we used weights pretrained on the ImageNet dataset,[19] made available by the Pytorch framework.[20] Our network optimizes Equation (1) from Jalaboi et al[17]:

$$L = \lambda_D L_D + \lambda_C L_C,$$

where L_D is the categorical cross-entropy loss for the image quality label

$$L_D = -\frac{1}{ND}\sum_{i=1}^{N}\sum_{j=1}^{D} y_{i,d} log\ \hat{y}_{i,d},$$

and L_C is the binary cross-entropy loss for poor image quality explanations

$$L_C = -\frac{1}{ND}\sum_{i=1}^{N}\sum_{j=1}^{D}\left[log\ \hat{z}_{i,z} + (1 - z_{i,c}) log(1 - z_{i,c})\right].$$

Table 2. Distribution of Poor Image Quality Explanations over the Training and Test Sets, Alongside the Pairwise Inter-Rater Agreement for Each Explanation, Calculated as the Pairwise F1-Score

REASON	TRAIN IMAGE COUNT	TEST IMAGE COUNT	PAIRWISE TRAIN F1	PAIRWISE TEST F1
Bad framing	1,947	982	0.26 ± 0.18	0.24 ± 0.15
Bad light	5,144	2,481	0.63 ± 0.07	0.65 ± 0.08
Blurry	5,499	2,640	0.81 ± 0.05	0.83 ± 0.06
Low resolution	3,965	1,907	0.33 ± 0.14	0.32 ± 0.14
Too far away	936	497	0.48 ± 0.16	0.51 ± 0.30
Mean	4372.75	2126.75	0.63 ± 0.15	0.64 ± 0.18

We set $\lambda_D = 1.0$ and $\lambda_C = 5.0$. To address the imbalance in image quality labels, we used class weighted training. Weights were set inversely proportional to frequency in training set, as follows:

$$w_c = min\left(\frac{n_{max}}{n_c},\ 10.0\right),$$

where w_c is the weight associated with each sample in class c, n_c is the number of samples in class c, and n_{max} is the number of samples in the most common class. Class weights were clipped to 10.0 to avoid overfitting on small classes. This process resulted in 1.0, 10.0, 4.49, and 3.70 as weights for *lesion*, *no skin*, *healthy skin*, and *poor quality*, respectively. The network was trained for 39 epochs with the AdamW optimizer,[21] cosine annealing with warm restarts,[22] 64 U in each linear block, and 0.2 dropout. Five runs with identical hyperparameters were performed to estimate the standard deviation between training runs.

Results

Table 3 provides the image quality assessment performance, whereas *Table 4* provides the performance on each poor image quality explanation. The F1-scores for *healthy skin* and *poor quality* are within standard deviation of the inter-rater agreement, whereas for *lesion* and *no skin* the performance is slightly lower. The lower performance on *no skin* may be explained by the limited training data available. For poor image quality explanations, all F1-scores except for *blurry* are within standard deviation of the mean inter-rater agreement. The high specificity visible in both image quality assessment and in poor image quality explanation suggests that deploying this network on patient phones would not negatively impact the patient experience by rejecting high-quality images.

Figure 4 provides the Grad-CAM attention maps for each poor image quality explanation detected in a *blurry* image. ImageQX correctly detected *blurry* as one of the poor image quality explanations, focusing almost entirely on the skin area and paying more attention to the lesion. Two other explanations were also marked as present: *bad light* with a focus on a slightly shaded part of the arm, and *low resolution* that highlights the edges of the hand and a part of the background.

Table 3. ImageQX Performance on Image Quality Assessment over Five Training Runs (Mean ± Standard Deviation)

CLASS	SENSITIVITY	SPECIFICITY	F1-SCORE
Lesion	0.84 ± 0.03	0.78 ± 0.04	**0.82 ± 0.00**
No skin	0.76 ± 0.05	0.99 ± 0.00	0.74 ± 0.02
Healthy skin	0.61 ± 0.09	0.90 ± 0.02	**0.63 ± 0.04**
Poor quality	0.71 ± 0.02	0.93 ± 0.00	**0.74 ± 0.01**
Mean	0.73 ± 0.01	0.90 ± 0.01	**0.73 ± 0.01**

F1-scores in bold show the assessments where ImageQX reaches expert-level performance.

Table 4. ImageQX Performance on Poor Image Quality Explanation Performance over Five Training Runs (Mean ± Standard Deviation)

REASON	SENSITIVITY	SPECIFICITY	F1-SCORE
Bad framing	0.31 ± 0.01	0.96 ± 0.00	**0.37 ± 0.01**
Bad light	0.58 ± 0.02	0.90 ± 0.01	**0.61 ± 0.00**
Blurry	0.60 ± 0.02	0.95 ± 0.00	0.70 ± 0.01
Low resolution	0.47 ± 0.02	0.92 ± 0.01	**0.52 ± 0.01**
Too far away	0.35 ± 0.02	0.98 ± 0.00	**0.42 ± 0.02**
Mean	0.39 ± 0.01	0.95 ± 0.00	**0.45 ± 0.01**

F1-scores in bold show the explanations where ImageQX reaches expert-level performance.

Discussion

Our data-labeling process confirms the previously reported findings that poor image quality is a significant issue in teledermatology—around 20% of the images collected through the mobile application were labeled as poor quality by dermatologists. Dermatologists have low levels of agreement on which images are poor quality, with inter-rater F1-scores of 0.62 ± 0.08. Explaining what makes an image poor quality is a difficult task, with inter-rater F1-scores varying between 0.26 and 0.81. Part of the disagreement can be ascribed to personal preference and level of experience with teledermatology, as some dermatologists tend to reject a larger proportion of images than others.

ImageQX reaches dermatologist-level performance on assessing the image quality on all quality assessment classes except for *no skin*. One reason for this lapse may be the low amount of training data for images with no skin. For poor image quality explanations, ImageQX obtains F1-scores within a standard deviation of the inter-rater agreement for all explanations except *blurry*.

Within a real world use case, the high specificity on both the image quality assessment and poor image quality explanation suggests that the image retake burden placed on the users would be low—only truly low-quality or irrelevant images would be flagged for retake. A low percentage of *poor quality*, *no skin*, or *healthy skin* images are likely to be seen by

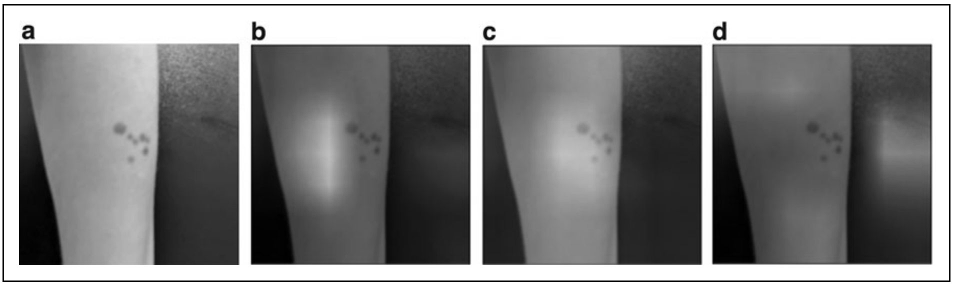


Fig. 4. Grad-CAM attention maps for the *blurry* test image introduced in *Figure 3*. The image was correctly classified as *poor quality*. **(a)** Original *blurry* image. **(b)** Grad-CAM attention map for *bad light*. **(c)** Grad-CAM attention map for *blurry*. **(d)** Grad-CAM attention map for *low resolution*. When predicting *bad light*, ImageQX focuses on a slightly shaded part of the arm, whereas for *blurry* it highlights the lesion and its surrounding area. The *low-resolution* prediction is based on the edges of the arm and the background. Image courtesy of the authors. Grad-CAM, gradient-based class activation map.

dermatologists. Poor image quality explanations also show high specificity, indicating that, if given proper guidance on how to fix each issue, users would find them useful in their retake attempt. By changing the threshold for poor quality image detection or for the image quality explanations we can further reduce the poor quality images sent to the dermatologists. Such an intervention should be carried out after thorough testing with both patients and dermatologists to ensure that we identify the ideal balance between asking patients to retake the images without being too disruptive.

A Grad-CAM analysis of the poor image quality explanations on an example image shows that ImageQX mostly bases its decisions on relevant areas. The *blurry* attention map is focused on the blurry lesion, whereas *bad light* concentrates on a slightly shaded area to the left of the lesion. *Low resolution* illustrates the debugging capabilities of Grad-CAMs: ImageQX bases its assessment primarily on the background rather than the original image. If these attention maps were to be presented to users alongside the explanations, they could help focus the users' attention to which sections of the image require improvement. For example, the Grad-CAM map for *blurry* suggests that the users should focus on the lesion instead of ensuring that the background is not blurred.

These findings open up several exploration avenues. First, by adding more nonskin images from publicly available datasets we could improve the *no skin* performance. This dataset addition requires the data to be from the same distribution, that is, smartphone images, to avoid in-class domain shift. Second, to more accurately model the uncertainty inherent in the image quality assessment task, we could train ImageQX using soft labels. Third, we believe that by introducing a skin segmentation network as preprocessing we would avoid misclassifications because of ImageQX focusing on the background. One drawback of this approach is the failure case of the segmentation network: if the segmentation removes the areas containing skin, the image quality assessment classifier is bound to fail. Finally, we would like to perform a usability study to quantify the impact an on-device image quality assessment network would have on the time to diagnosis and treatment in a teledermatology setting. Such a study would require an in-depth analysis of how to best communicate the image quality assessments and explanations to the patients.

Conclusions

Our work on ImageQX introduced several elements of novelty. First, we quantified the dermatologist levels of agreement on what constitutes a high-quality image for a teledermatological consultation and their reasoning when tagging images as low quality. Second, we introduced ImageQX, an expert-level image quality assessor that explains its reasoning for marking an image as poor quality. The added explainability component aims to facilitate the patient understanding on how to improve images. Moreover, with a size of only 15 MB, ImageQX can be easily packaged and deployed in a teledermatology mobile application, and thus incorporated as a step between users taking photos and sending them. Having such a network integrated in the application during the data collection step of this study would have prevented 1,819 *poor quality* or *no skin* images from being sent for assessment to the dermatologists. In the future, we will perform a validation study to quantify the impact of introducing such a method within a consumer facing teledermatology setting.

Our solution offers an improvement to the current consumer facing teledermatology flow by increasing the likelihood that patients send better photos, decreasing the time spent by dermatologists on diagnosing a single patient, and reducing the time needed to arrive at a diagnosis and a treatment for patients.

Acknowledgments

The authors thank the work of the 12 board certified dermatologists for their work on labeling the dataset, and the entire Imagine team for enabling the data collection step.

Authors' Contributions

R.J.: Conceptualization (equal), methodology, software, validation, formal analysis, funding acquisition, writing—original draft. O.W.: Conceptualization (equal), writing—review and editing (supporting), supervision (supporting). A.G.: Conceptualization (equal), resources, writing—review and editing (lead), supervision (lead), project administration.

Disclosure Statement

No competing financial interests exist.

Funding Information

R.J.'s work was supported in part by the Danish Innovation Fund under Grant 0153-00154A. O.W.'s work was funded in part by the Novo Nordisk Foundation through the Center for Basic Machine Learning Research in Life Science (NNF20OC0062606). O.W. acknowledges support from the Pioneer Centre for AI, DNRF grant number P1.

REFERENCES

1. Yeboah CB, Harvey N, Krishnan R, et al. The impact of COVID-19 on teledermatology: A review. Dermatol Clin 2021;39(4):599–608.
2. Landow SM, Mateus A, Korgavkar K, et al. Teledermatology: Key factors associated with reducing face-to-face dermatology visits. J Am Acad Dermatol 2014;71(3):570–576.
3. Haque W, Chandy R, Ahmadzada M, et al. Teledermatology after COVID-19: Key challenges ahead. Dermatol Online J 2021;27(4):13030.
4. Pasquali P, Sonthalia S, Moreno-Ramirez D, et al. Teledermatology and its current perspective. Indian Dermatol Online J 2020;11(1):12–20.
5. Vodrahalli K, Daneshjou R, Novoa RA, et al. TrueImage: a machine learning algorithm to improve the quality of telehealth photos. In: BIOCOMPUTING 2021: Proceedings of the Pacific Symposium. World Scientific Publishing Company: Singapore; 2020; pp. 220–231.
6. Lee D, Choi S, Kim HJ. Performance evaluation of image denoising developed using convolutional denoising autoencoders in chest radiography. Nucl Instrum Methods Phys Res A Accel Spectrom Detect Assoc Equip 2018;884:97–104.
7. Kim J, Lee S. Deep learning of human visual sensitivity in image quality assessment framework. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. IEEE: USA; 2017; pp. 1676–1684.
8. Bianco S, Celona L, Napoletano P, et al. On the use of deep learning for blind image quality assessment. Signal Image Video Process 2018;12(2):355–362.
9. Ghadiyaram D, Bovik AC. Crowdsourced study of subjective image quality. In: 2014 48th Asilomar Conference on Signals, Systems and Computers. IEEE: USA; 2014; pp. 84–88.
10. Madhusudana PC, Birkbeck N, Wang Y, et al. Image quality assessment using contrastive learning. IEEE Trans Image Process 2022;31:4149–4161.
11. Goodman B, Flaxman S. European Union regulations on algorithmic decision-making and a "right to explanation." AI Mag 2017;38(3):50–57.
12. Kelly CJ, Karthikesalingam A, Suleyman M, et al. Key challenges for delivering clinical impact with artificial intelligence. BMC Med 2019;17(1):1–9.
13. Topol EJ. High-performance medicine: The convergence of human and artificial intelligence. Nat Med 2019;25(1):44–56.
14. Selvaraju RR, Cogswell M, Das A, et al. Grad-cam: Visual explanations from deep networks via gradient-based localization. In: Proceedings of the IEEE International Conference on Computer Vision. IEEE: USA; 2017; pp. 618–626.
15. LEO Innovation Lab. Imagine[Internet]. Copenhagen (DK): LEO Innovation Lab; 2017 [updated 2022; cited 2022 Sep 15 Available from: https://getimagine.io [Last accessed: September 15, 2022].
16. World Health Organization. The ICD-10 Classification of Mental and Behavioural Disorders: Clinical Descriptions and Diagnostic Guidelines. World Health Organization: Geneva; 1992.
17. Jalaboi R, Faye F, Orbes-Arteaga M, et al. DermX: An end-to-end framework for explainable automated dermatological diagnosis. Med Image Anal 2023;83: 102647.
18. Tan M, Le Q. Efficientnet: Rethinking model scaling for convolutional neural networks. In: International Conference on Machine Learning. PMLR: USA; 2019; pp. 6105–6114.
19. Deng J, Dong W, Socher R, et al. Imagenet: A large-scale hierarchical image database. In: 2009 IEEE Conference on Computer Vision and Pattern Recognition. IEEE: USA; 2009; pp. 248–255.
20. Paszke A, Gross S, Massa F, et al. Pytorch: An imperative style, high-performance deep learning library. Adv Neural Inform Process Syst 2019;32: 8024–8035.
21. Loshchilov I, Hutter F. Decoupled weight decay regularization. arXiv preprint arXiv:1711.05101 (2017).
22. Loshchilov I, Hutter F. Sgdr: Stochastic gradient descent with warm restarts. arXiv preprint arXiv:1608.03983 2016.

Address correspondence to:
Raluca Jalaboi, MS
Department of Applied Mathematics and Computer Science
Technical University of Denmark
DK-2800 Kongens Lyngby
Denmark

E-mail: rjal@dtu.dk

Received: September 15, 2022
Revised: November 18, 2022
Accepted: November 18, 2022
Online Publication Date: February 2, 2023

Bibliography

Raluca Jalaboi, Frederik Faye, Mauricio Orbes-Arteaga, Dan Jørgensen, Ole Winther, and Alfiia Galimzianova. Dermx: An end-to-end framework for explainable automated dermatological diagnosis. *Medical Image Analysis*, 83:102647, 2023a.

Raluca Jalaboi, Ole Winther, and Alfiia Galimzianova. Dermatological diagnosis explainability benchmark for convolutional neural networks. *arXiv preprint arXiv:2302.12084*, 2023b.

Raluca Jalaboi, Ole Winther, and Alfiia Galimzianova. Explainable image quality assessments in teledermatological photography. *Telemedicine and e-Health*, 2023c.

Raluca Jalaboi, Mauricio Orbes Arteaga, Dan Richter Jørgensen, Ionela Manole, Oana Ionescu Bozdog, Andrei Chiriac, Ole Winther, and Alfiia Galimzianova. Explainability of convolutional neural networks for dermatological diagnosis. *iproc*, 6(1):e35437, Dec 2021. ISSN 2369-6893. doi: 10.2196/35437. URL `https://www.iproc.org/2021/1/e35437`.

Kenneth Thomsen, Raluca Jalaboi, Ole Winther, Hans Bredsted Lomholt, Henrik Lorentzen, Trine Høgsberg, Henrik Egekvist, Lene Hedelund, Sofie Jørgensen, Sanne Frost, and Lars Iversen. Physician level assessment of hirsute women and eligibility for laser treatment with deep learning. Unpublished, 2023.

Roderick J Hay, Nicole E Johns, Hywel C Williams, Ian W Bolliger, Robert P Dellavalle, David J Margolis, Robin Marks, Luigi Naldi, Martin A Weinstock, Sarah K Wulf, et al. The global burden of skin disease in 2010: an analysis of the prevalence and impact of skin conditions. *Journal of Investigative Dermatology*, 134(6):1527–1534, 2014.

Lewin Group. The burden of skin diseases, 2005.

Daniel G Federman, John Concato, and Robert S Kirsner. Comparison of dermatologic diagnoses by primary care practitioners and dermatologists: a review of the literature. *Archives of family medicine*, 8(2):170, 1999.

Mohammad Shehab, Laith Abualigah, Qusai Shambour, Muhannad A Abu-Hashem, Mohd Khaled Yousef Shambour, Ahmed Izzat Alsalibi, and Amir H Gandomi.

Machine learning in medical applications: A review of state-of-the-art methods. *Computers in Biology and Medicine*, 145:105458, 2022.

Yann LeCun, Léon Bottou, Yoshua Bengio, and Patrick Haffner. Gradient-based learning applied to document recognition. *Proceedings of the IEEE*, 86(11):2278–2324, 1998.

Isabella Castiglioni, Leonardo Rundo, Marina Codari, Giovanni Di Leo, Christian Salvatore, Matteo Interlenghi, Francesca Gallivanone, Andrea Cozzi, Natascha Claudia D'Amico, and Francesco Sardanelli. Ai applications to medical images: From machine learning to deep learning. *Physica Medica*, 83:9–24, 2021.

Andre Esteva, Brett Kuprel, Roberto A Novoa, Justin Ko, Susan M Swetter, Helen M Blau, and Sebastian Thrun. Dermatologist-level classification of skin cancer with deep neural networks. *nature*, 542(7639):115–118, 2017.

Kenneth Thomsen, Lars Iversen, Therese Louise Titlestad, and Ole Winther. Systematic review of machine learning for diagnosis and prognosis in dermatology. *Journal of Dermatological Treatment*, 31(5):496–510, 2020.

Philipp Tschandl, Cliff Rosendahl, and Harald Kittler. The HAM10000 dataset, a large collection of multi-source dermatoscopic images of common pigmented skin lesions. *Scientific Data*, 5(1):1–9, 2018.

Philipp Tschandl, Christoph Rinner, Zoe Apalla, Giuseppe Argenziano, Noel Codella, Allan Halpern, Monika Janda, Aimilios Lallas, Caterina Longo, Josep Malvehy, et al. Human–computer collaboration for skin cancer recognition. *Nature Medicine*, 26(8):1229–1234, 2020.

Abhijit Guha Roy, Jie Ren, Shekoofeh Azizi, Aaron Loh, Vivek Natarajan, Basil Mustafa, Nick Pawlowski, Jan Freyberg, Yuan Liu, Zach 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.

Christopher J Kelly, Alan Karthikesalingam, Mustafa Suleyman, Greg Corrado, and Dominic King. Key challenges for delivering clinical impact with artificial intelligence. *BMC Medicine*, 17(1):1–9, 2019.

Christoph Molnar. *Interpretable machine learning*. Lulu. com, 2020.

Di Jin, Elena Sergeeva, Wei-Hung Weng, Geeticka Chauhan, and Peter Szolovits. Explainable deep learning in healthcare: A methodological survey from an attribution view. *WIREs Mechanisms of Disease*, 14(3):e1548, 2022.

Bas HM Van der Velden, Hugo J Kuijf, Kenneth GA Gilhuijs, and Max A Viergever. Explainable artificial intelligence (xai) in deep learning-based medical image analysis. *Medical Image Analysis*, page 102470, 2022.

Pranav Rajpurkar, Jeremy Irvin, Robyn L Ball, Kaylie Zhu, Brandon Yang, Hershel Mehta, Tony Duan, Daisy Ding, Aarti Bagul, Curtis P Langlotz, et al. Deep learning for chest radiograph diagnosis: A retrospective comparison of the chexnext algorithm to practicing radiologists. *PLoS medicine*, 15(11):e1002686, 2018.

Cristina González-Gonzalo, Eric F Thee, Caroline CW Klaver, Aaron Y Lee, Reinier O Schlingemann, Adnan Tufail, Frank Verbraak, and Clara I Sánchez. Trustworthy ai: Closing the gap between development and integration of ai systems in ophthalmic practice. *Progress in retinal and eye research*, 90:101034, 2022.

James E Fitzpatrick and Whitney A High. *Urgent Care Dermatology: Symptom-Based Diagnosis E-Book*. Elsevier Health Sciences, 2017.

Amanda Oakley. *Dermatology Made Easy*. Scion Publishing Ltd, The Old Hayloft, Vantage Business Park, Bloxham Road, Banbury OX16 9UX, UK, 2017.

Alexander Nast, Chris E M Griffiths, Roderick Hay, Wolfram Sterry, and Jean L Bolognia. The 2016 International League of Dermatological Societies' revised glossary for the description of cutaneous lesions. *British Journal of Dermatology*, 174 (6):1351–1358, 2016.

Tim Miller. Explanation in artificial intelligence: Insights from the social sciences. *Artificial intelligence*, 267:1–38, 2019.

Tania Lombrozo. The structure and function of explanations. *Trends in cognitive sciences*, 10(10):464–470, 2006.

Robert James Hankinson. *Cause and explanation in ancient Greek thought*. Clarendon Press, 1998.

Herbert P Grice. Logic and conversation. In *Speech acts*, pages 41–58. Brill, 1975.

Andreas Holzinger, Chris Biemann, Constantinos S Pattichis, and Douglas B Kell. What do we need to build explainable ai systems for the medical domain? *arXiv preprint arXiv:1712.09923*, 2017.

Maartje MA De Graaf and Bertram F Malle. How people explain action (and autonomous intelligent systems should too). In *2017 AAAI Fall Symposium Series*, 2017.

Jay DeYoung, Sarthak Jain, Nazneen Fatema Rajani, Eric Lehman, Caiming Xiong, Richard Socher, and Byron C Wallace. Eraser: A benchmark to evaluate rationalized nlp models. *arXiv preprint arXiv:1911.03429*, 2019.

Bryce Goodman and Seth Flaxman. European union regulations on algorithmic decision-making and a "right to explanation". *AI magazine*, 38(3):50–57, 2017.

Sebastian Lapuschkin, Stephan Wäldchen, Alexander Binder, Grégoire Montavon, Wojciech Samek, and Klaus-Robert Müller. Unmasking clever hans predictors and assessing what machines really learn. *Nature communications*, 10(1):1096, 2019.

Julia Dressel and Hany Farid. The accuracy, fairness, and limits of predicting recidivism. *Science advances*, 4(1):eaao5580, 2018.

Julia K Winkler, Christine Fink, Ferdinand Toberer, Alexander Enk, Teresa Deinlein, Rainer Hofmann-Wellenhof, Luc Thomas, Aimilios Lallas, Andreas Blum, Wilhelm 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.

Oriol Vinyals, Alexander Toshev, Samy Bengio, and Dumitru Erhan. Show and tell: A neural image caption generator. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 3156–3164, 2015.

Chaofan Chen, Oscar Li, Daniel Tao, Alina Barnett, Cynthia Rudin, and Jonathan K Su. This looks like that: deep learning for interpretable image recognition. *Advances in neural information processing systems*, 32, 2019.

Karen Simonyan, Andrea Vedaldi, and Andrew Zisserman. Deep inside convolutional networks: Visualising image classification models and saliency maps. *arXiv preprint arXiv:1312.6034*, 2013.

Ramprasaath R Selvaraju, Michael Cogswell, Abhishek Das, Ramakrishna Vedantam, Devi Parikh, and Dhruv 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.

Amitojdeep Singh, Sourya Sengupta, and Vasudevan Lakshminarayanan. Explainable deep learning models in medical image analysis. *Journal of Imaging*, 6(6):52, 2020.

Mukund Sundararajan, Ankur Taly, and Qiqi Yan. Axiomatic attribution for deep networks. In *International conference on machine learning*, pages 3319–3328. PMLR, 2017.

Matthew D Zeiler and Rob Fergus. Visualizing and understanding convolutional networks. In *Computer Vision–ECCV 2014: 13th European Conference, Zurich, Switzerland, September 6-12, 2014, Proceedings, Part I 13*, pages 818–833. Springer, 2014.

Marco Tulio Ribeiro, Sameer Singh, and Carlos Guestrin. ”Why should i trust you?” Explaining the predictions of any classifier. In *Proceedings of the 22nd ACM SIGKDD international conference on knowledge discovery and data mining*, pages 1135–1144, 2016.

Hristina Uzunova, Jan Ehrhardt, Timo Kepp, and Heinz Handels. Interpretable explanations of black box classifiers applied on medical images by meaningful perturbations using variational autoencoders. In *Medical Imaging 2019: Image Processing*, volume 10949, pages 264–271. SPIE, 2019.

Catarina Barata, M Emre Celebi, and Jorge S Marques. Explainable skin lesion diagnosis using taxonomies. *Pattern Recognition*, 110:107413, 2021.

Manxi Lin, Aasa Feragen, Zahra Bashir, Martin Grønnebæk Tolsgaard, and Anders Nymark Christensen. I saw, i conceived, i concluded: Progressive concepts as bottlenecks. *arXiv preprint arXiv:2211.10630*, 2022.

Srishti Gautam, Ahcene Boubekki, Stine Hansen, Suaiba Salahuddin, Robert Jenssen, Marina Höhne, and Michael Kampffmeyer. Protovae: A trustworthy self-explainable prototypical variational model. *Advances in Neural Information Processing Systems*, 35:17940–17952, 2022.

Jay DeYoung, Sarthak Jain, Nazneen Fatema Rajani, Eric Lehman, Caiming Xiong, Richard Socher, and Byron C Wallace. Eraser: A benchmark to evaluate rationalized nlp models. *Transactions of the Association for Computational Linguistics*, 2020.

Tim Rädsch, Annika Reinke, Vivienn Weru, Minu D Tizabi, Nicholas Schreck, A Emre Kavur, Bünyamin Pekdemir, Tobias Roß, Annette Kopp-Schneider, and Lena Maier-Hein. Labelling instructions matter in biomedical image analysis. *Nature Machine Intelligence*, 5(3):273–283, 2023.

Xiaoxiao Sun, Jufeng Yang, Ming Sun, and Kai Wang. A benchmark for automatic visual classification of clinical skin disease images. In *European Conference on Computer Vision*, pages 206–222. Springer, 2016.

Noel CF Codella, David Gutman, M Emre Celebi, Brian Helba, Michael A Marchetti, Stephen W Dusza, Aadi Kalloo, Konstantinos Liopyris, Nabin Mishra, Harald Kittler, 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)*, pages 168–172. IEEE, 2018.

DermNetNZ. Dermnetnz. `https://dermnetnz.org/`, 2021. Accessed: 2021-04-01.

G Reshma, Chiai Al-Atroshi, V Kumar Nassa, B Geetha, Gurram Sunitha, Mohammad Gouse Galety, and S Neelakandan. Deep learning-based skin lesion diagnosis model using dermoscopic images. *Intelligent Automation and Soft Computing*, 31 (1):621–634, 2022.

Jianpeng Zhang, Yutong Xie, Yong Xia, and Chunhua Shen. Attention residual learning for skin lesion classification. *IEEE transactions on medical imaging*, 38(9): 2092–2103, 2019.

Seung Seog Han, Ilwoo Park, Sung Eun Chang, Woohyung Lim, Myoung Shin Kim, Gyeong Hun Park, Je Byeong Chae, Chang Hun Huh, and Jung-Im Na. Augmented intelligence dermatology: deep neural networks empower medical professionals in diagnosing skin cancer and predicting treatment options for 134 skin disorders. *Journal of Investigative Dermatology*, 140(9):1753–1761, 2020.

Haijing Wu, Heng Yin, Haipeng Chen, Moyuan Sun, Xiaoqing Liu, Yizhou Yu, Yang Tang, Hai Long, Bo Zhang, Jing Zhang, et al. A deep learning-based smartphone platform for cutaneous lupus erythematosus classification assistance: Simplifying the diagnosis of complicated diseases. *Journal of the American Academy of Dermatology*, 85(3):792–793, 2021.

Pushkar Aggarwal and Francis A Papay. Artificial intelligence image recognition of melanoma and basal cell carcinoma in racially diverse populations. *Journal of Dermatological Treatment*, 33(4):2257–2262, 2022.

Wei Ba, Huan Wu, Wei W Chen, Shu H Wang, Zi Y Zhang, Xuan J Wei, Wen J Wang, Lei Yang, Dong M Zhou, Yi X Zhuang, et al. Convolutional neural network assistance significantly improves dermatologists' diagnosis of cutaneous tumours using clinical images. *European Journal of Cancer*, 169:156–165, 2022.

Martin Holm Jensen, Dan Richter Jørgensen, Raluca Jalaboi, Mads Eiler Hansen, and Martin Aastrup Olsen. Improving uncertainty estimation in convolutional neural networks using inter-rater agreement. In *Medical Image Computing and Computer Assisted Intervention–MICCAI 2019: 22nd International Conference, Shenzhen, China, October 13–17, 2019, Proceedings, Part IV 22*, pages 540–548. Springer, 2019.

YPH Chin, ZY Hou, MY Lee, HM Chu, HH Wang, YT Lin, A Gittin, SC Chien, PA Nguyen, LC Li, et al. A patient-oriented, general-practitioner-level, deep-learning-based cutaneous pigmented lesion risk classifier on a smartphone. *The British Journal of Dermatology*, 182(6):1498–1500, 2020.

Sk Imran Hossain, Jocelyn de Goër de Herve, Md Shahriar Hassan, Delphine Martineau, Evelina Petrosyan, Violaine Corbin, Jean Beytout, Isabelle Lebert, Jonas Durand, Irene Carravieri, et al. Exploring convolutional neural networks with transfer learning for diagnosing lyme disease from skin lesion images. *Computer Methods and Programs in Biomedicine*, 215:106624, 2022.

Seonguk Min, Hyoun-joong Kong, Chiyul Yoon, Hee Chan Kim, and Dae Hun Suh. Development and evaluation of an automatic acne lesion detection program using digital image processing. *Skin Research and Technology*, 19(1):e423–e432, 2013.

Yading Yuan, Ming Chao, and Yeh-Chi Lo. Automatic skin lesion segmentation using deep fully convolutional networks with jaccard distance. *IEEE Transactions on Medical Imaging*, 36(9):1876–1886, 2017.

Ramsha Baig, Maryam Bibi, Anmol Hamid, Sumaira Kausar, and Shahzad Khalid. Deep learning approaches towards skin lesion segmentation and classification from dermoscopic images-a review. *Current Medical Imaging*, 16(5):513–533, 2020.

Huisi Wu, Shihuai Chen, Guilian Chen, Wei Wang, Baiying Lei, and Zhenkun Wen. Fat-net: Feature adaptive transformers for automated skin lesion segmentation. *Medical Image Analysis*, 76:102327, 2022.

Sophie Seité, Amir Khammari, Michael Benzaquen, Dominique Moyal, and Brigitte Dréno. Development and accuracy of an artificial intelligence algorithm for acne grading from smartphone photographs. *Experimental dermatology*, 28(11):1252–1257, 2019.

A Munthuli, J Intanai, P Tossanuch, P Pooprasert, P Ingpochai, S Boonyasatian, K Kittithammo, P Thammarach, T Boonmak, S Khaengthanyakan, et al. Extravasation screening and severity prediction from skin lesion image using deep neural networks. In *2022 44th Annual International Conference of the IEEE Engineering in Medicine & Biology Society (EMBC)*, pages 1827–1833. IEEE, 2022.

Hang Zhang and Tianyi Ma. Acne detection by ensemble neural networks. *Sensors*, 22(18):6828, 2022.

Jongyoo Kim and Sanghoon Lee. Deep learning of human visual sensitivity in image quality assessment framework. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 1676–1684, 2017.

Simone Bianco, Luigi Celona, Paolo Napoletano, and Raimondo Schettini. On the use of deep learning for blind image quality assessment. *Signal, Image and Video Processing*, 12(2):355–362, 2018.

Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. Imagenet: A large-scale hierarchical image database. In *2009 IEEE conference on computer vision and pattern recognition*, pages 248–255. Ieee, 2009.

Philippe M Burlina, Neil J Joshi, Elise Ng, Seth D Billings, Alison W Rebman, and John N Aucott. Automated detection of erythema migrans and other confounding skin lesions via deep learning. *Computers in biology and medicine*, 105:151–156, 2019.

Veysel Harun Sahin, Ismail Oztel, and Gozde Yolcu Oztel. Human monkeypox classification from skin lesion images with deep pre-trained network using mobile application. *Journal of Medical Systems*, 46(11):79, 2022.

Ayush Jain, David Way, Vishakha Gupta, Yi Gao, Guilherme de Oliveira Marinho, Jay Hartford, Rory Sayres, Kimberly Kanada, Clara Eng, Kunal 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.

Sandesh Shah, Rashi Pangti, Lavina Rajput, Jyoti Mathur, Vikas Chouhan, Sharad Kumar, Dhwani Dholakia, Vishal Gupta, Vinod K Sharma, and Somesh Gupta. Comparison of performance of a deep learning-based mobile application with non-dermatologist physicians in the diagnosis of common skin diseases. *International journal of dermatology*, 60(9):e365–e366, 2021.

F Flament, L Jacquet, C Ye, D Amar, D Kerob, R Jiang, Y Zhang, C Kroely, C Delaunay, and T Passeron. Artificial intelligence analysis of over half a million european and chinese women reveals striking differences in the facial skin ageing process. *Journal of the European Academy of Dermatology and Venereology*, 36(7):1136–1142, 2022.

Christine Fink, A Blum, T Buhl, C Mitteldorf, R Hofmann-Wellenhof, T Deinlein, W Stolz, Lukas Trennheuser, Christiane Cussigh, David Deltgen, et al. Diagnostic performance of a deep learning convolutional neural network in the differentiation of combined naevi and melanomas. *Journal of the European Academy of Dermatology and Venereology*, 34(6):1355–1361, 2020.

Yuan Liu, Ayush Jain, Clara Eng, David H Way, Kang Lee, Peggy Bui, Kimberly Kanada, Guilherme de Oliveira Marinho, Jessica Gallegos, Sara Gabriele, et al. A deep learning system for differential diagnosis of skin diseases. *Nature medicine*, 26 (6):900–908, 2020.

Hasib Zunair and A Ben Hamza. Melanoma detection using adversarial training and deep transfer learning. *Physics in Medicine & Biology*, 65(13):135005, 2020.

Yutong Xie, Jianpeng Zhang, Yong Xia, and Chunhua Shen. A mutual bootstrapping model for automated skin lesion segmentation and classification. *IEEE transactions on medical imaging*, 39(7):2482–2493, 2020.

Masaya Tanaka, Atsushi Saito, Kosuke Shido, Yasuhiro Fujisawa, Kenshi Yamasaki, Manabu Fujimoto, Kohei Murao, Youichirou Ninomiya, Shin'ichi Satoh, and Akinobu Shimizu. Classification of large-scale image database of various skin diseases using deep learning. *International Journal of Computer Assisted Radiology and Surgery*, 16:1875–1887, 2021.

Yiqi Yan, Jeremy Kawahara, and Ghassan Hamarneh. Melanoma recognition via visual attention. In *Information Processing in Medical Imaging: 26th International Conference, IPMI 2019, Hong Kong, China, June 2–7, 2019, Proceedings 26*, pages 793–804. Springer, 2019.

V7-Labs. Darwin v7-labs. `https://darwin.v7labs.com`, 2021. Accessed: 2021-05-01.

Paola Pasquali, Sidharth Sonthalia, David Moreno-Ramirez, Pooram Sharma, Mahima Agrawal, Somesh Gupta, Dinesh Kumar, and Dharmendra Arora. Teledermatology and its current perspective. *Indian dermatology online journal*, 11(1): 12, 2020.