
DermX: a Dermatological Diagnosis Explainability Dataset

Raluca Jalaboi^{1, 2}, Mauricio Orbes-Arteaga², Dan Richter Jørgensen², Ionela Manole², Oana Ionescu Bozdog², Andrei Chiriac², Ole Winther^{1, 3, 4}, Alfiia Galimzianova²

¹Section for Cognitive Systems, Technical University of Denmark

²Omhu, Denmark

³Bioinformatics Centre, Department of Biology, University of Copenhagen

⁴Centre for Genomic Medicine, Rigshospitalet, Copenhagen University Hospital

{rjal, olwi}@dtu.dk, {raluca, mauricio.orbes, dan, ionela.manole, oana.bozdog, andrei.chiriac, alfiia}@omhu.com

Abstract

In this paper, we introduce DermX: a novel dermatological diagnosis and explanations dataset annotated by eight board-certified dermatologists. To date, public datasets for dermatological applications have been limited to diagnosis and lesion segmentation, while validation of dermatological explainability has been limited to visual inspection. As such, this work is a first release of a dataset providing gold standard explanations for dermatological diagnosis to enable a quantitative evaluation of ConvNet explainability. DermX consists of 525 images sourced from two public datasets, DermNetNZ and SD-260, spanning six of the most prevalent skin conditions. Each image was enriched with diagnoses and diagnosis explanations by three dermatologists. Supporting explanations were collected as 15 non-localisable characteristics, 16 localisable characteristics, and 23 additional terms. Dermatologists manually segmented localisable characteristic and described them with additional terms. We showcase a possible use of our dataset by benchmarking the explainability of two ConvNet architectures, ResNet-50 and EfficientNet-B4, trained on an internal skin lesion dataset and tested on DermX. ConvNet visualisations are obtained through gradient-weighted class-activation map (Grad-CAM), a commonly used model visualisation technique. Our analysis reveals EfficientNet-B4 as the most explainable between the two. Thus, we prove that DermX can be used to objectively benchmark the explainability power of dermatological diagnosis models. The dataset is available at <https://github.com/ralucaj/dermx>.

1

Introduction

Convolutional neural models (ConvNets), the current state-of-the-art method for image analysis, are often criticised for being opaque in their decision mechanisms [1]. However, explainability is a crucial component in the adoption of machine learning systems in high-stakes applications, such as medical diagnosis. Dermatology in particular would highly benefit from automation, given the low diagnostic accuracy of general practitioners [2] and the scarcity of specialists [3, 4]. Deep learning methods to diagnose skin conditions exist [5–8], but their adoption by the medical system has been slow, partially due to their lack of explainability [9, 1, 10].

Different research groups proposed various explainability methods [11–13], but their use has been limited to visual inspection of the outputs to evaluate model performance. Such an approach is subjective and difficult to scale. Lesion segmentation masks offered by high quality dermatology

Table 1: Distribution of images over the public datasets. Initially, 100 images were randomly selected for each class, apart from viral warts and vitiligo where only 78 and 88 images were available. Some images were discarded during labelling, giving rise to the count shown below.

	Acne	Actinic keratosis	Psoriasis	Seborrhoeic dermatitis	Viral warts	Vitiligo
DermNetNZ	52	48	46	12	47	75
SD-260	47	43	51	66	27	11

32 datasets [14] can partially serve as a basis for objective measurement, although they were not collected
 33 to explain the diagnosis. However, this shortcoming becomes critical in diseases such as actinic
 34 keratosis, where the surrounding area is just as important for the diagnosis as the lesion itself [8].

35 We introduce DermX, a novel dermatological diagnosis explainability dataset that addresses the
 36 limitations of existing datasets by collecting dermatologist explanations for six skin diseases: acne,
 37 actinic keratosis, psoriasis, seborrhoeic dermatitis, viral warts, and vitiligo. Each image is diagnosed
 38 by three dermatologists and tagged with supporting characteristics [15] and their localisation.

39 To demonstrate the intended use of DermX, we benchmark two models trained to diagnose dermatolo-
 40 gical conditions. We employ gradient-weighted class-activation maps (Grad-CAM) [13], a deep
 41 learning visualisation technique commonly used to generate explanations, on ResNet-50 [16] and
 42 EfficientNet-B4 [17]. Then, we test how their explanations compare to the dermatologist maps.

43 The contributions of this paper are twofold:

- 44 1. We release a novel dermatological diagnosis explainability dataset with annotations from
 45 multiple expert raters;
- 46 2. We benchmark the explainability of two popular model architectures against a gold standard
 47 explainability dataset.

48 2 Dataset

49 DermX consists of 525 images of acne, actinic keratosis, psoriasis, seborrhoeic dermatitis, viral
 50 warts, or vitiligo patients. Eight board-certified dermatologists, with between 4 and 12 years of
 51 clinical experience, labelled the images with diagnoses and explanations supporting their diagnoses,
 52 in the form of both global tags and characteristic segmentations. The images were randomly selected
 53 from DermNetNZ [18] and SD-260 [19], and are available under the Creative Commons licence.
 54 Permission to use the data in this project was granted in writing by the owners of both datasets. The
 55 distribution of diseases is described in Table 1.

56 Our work involved several steps. First, we performed several experiments to define the target diseases
 57 and the nature of the explanations. Second, we defined the diagnosis and explanation ontology, as
 58 illustrated in Figure 1. Then, the labellers were allowed a short period of time to get accustomed to
 59 the annotation protocol and the labelling tool by evaluating images from an internal dataset. Finally,
 60 DermX images were selected and sent to the dermatologists for labelling.

61 2.1 Preliminary Investigation

62 Nine diseases were initially investigated: psoriasis, rosacea, vitiligo, seborrhoeic dermatitis, pityriasis
 63 rosea, viral warts, actinic keratosis, acne, and impetigo. These diseases were chosen based on preva-
 64 lence [20] and the expectation that they could be diagnosed only from images [21]. Dermatologists
 65 were asked to diagnose and explain their diagnosis in free-text for around 100 images. This step led
 66 to both the exclusion of rosacea, impetigo, and pityriasis rosea from future experiments due to the
 67 difficulty in diagnosing them in the absence of an anamnesis, and to the introduction of a structured
 68 ontology for the diagnosis explanations to avoid manual processing of typos and synonyms.

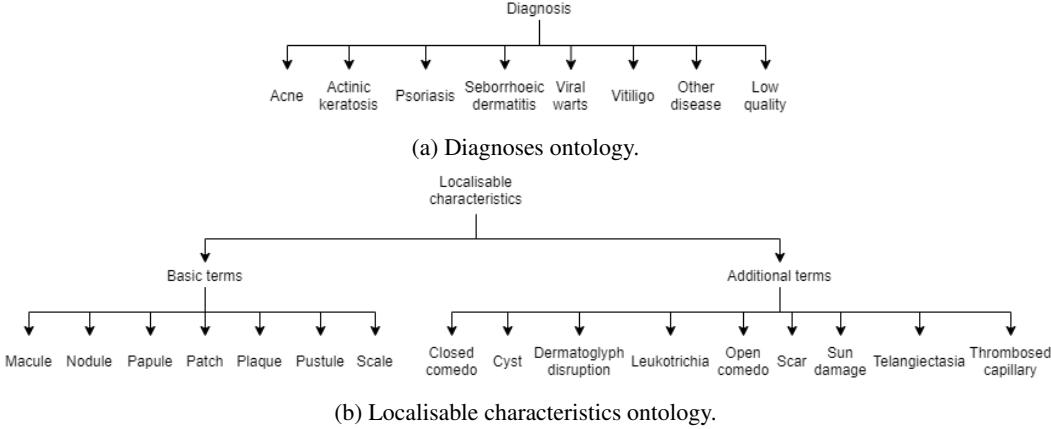


Figure 1: Ontology of the two main types of labels. The list of diagnoses (a) includes the six diseases and two discard options. Discard options could be chosen when images displayed another disease or when images were of low quality. Localisable characteristics (b) were tailored to the six diseases using medical resources [15, 21], and with the help of two senior dermatologists.

69 2.2 Diagnosis and Explanation Ontology

70 Preliminary investigations highlighted the importance of having a consistent explanation ontology.
 71 After analysing free-text explanations, they were formalised as an extended list of skin lesion charac-
 72 teristics [15]. The characteristics set was selected to sufficiently explain the six target diseases [21].
 73 With the help of two senior dermatologists, several other relevant characteristics were added.
 74 The resulting set of characteristics was split into non-localisable characteristics (e.g. age or sex),
 75 localisable characteristics (e.g. plaque or open comedo), and additional descriptive terms (e.g. red
 76 or well-circumscribed), according to the International League of Dermatological Societies (ILDS)
 77 classification [15]. To match state-of-the-art ConvNet explainability methods, we focus on diagnoses
 78 and localisable characteristics. Figure 1 illustrates the final DermX ontology, while more information
 79 about the other two types of labels is available in Appendix Figure 1.

80 2.3 Annotation Protocol

81 Dermatologists were first asked to diagnose the image, and then tag it with characteristics that explain
 82 their diagnosis. If the dermatologists were unable to evaluate the image due to poor quality, or if the
 83 image depicted a different disease than the target conditions, they had the option to discard it.
 84 Dermatologists could then select diagnosis-relevant non-localisable characteristics as global image
 85 tags. Afterwards, they could select and localise localisable characteristics. Dermatologists were
 86 instructed to highlight all relevant areas for each characteristic, and were only allowed to include
 87 irrelevant areas if separating them from the characteristic was too time consuming or difficult. In
 88 other words, they were instructed to favour sensitivity over specificity. Finally, basic terms (as defined
 89 in Figure 1b) could be enriched with additional descriptive terms when required for the diagnosis
 90 explanation. Once all tags and characteristics were added, the image could be marked as complete.
 91 After the ontology and annotation protocol were defined, all dermatologists underwent two rounds of
 92 on-boarding in Darwin, a browser-based labelling tool [22] (Appendix Figure 2).

93 2.4 Dataset Analysis

94 Once all evaluations were finished, we analysed the data focusing on dermatologist performance
 95 with regards to the gold standard diagnosis and their inter-rater agreement on both diagnoses and
 96 supporting characteristics. Figure 2 illustrates an image and its three annotations.
 97 From 1698 unique evaluations on 566 images, 411 evaluations were either tagged as other disease
 98 or as too low quality to evaluate. These 411 evaluations were removed from the dataset, leading to
 99 some images having fewer than three evaluations. Two evaluations tagged an image with multiple

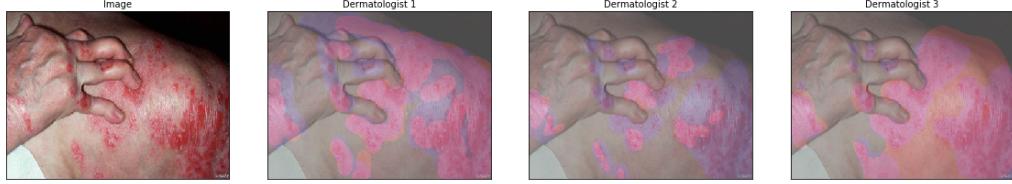


Figure 2: Example of a DermX image and its labels from three dermatologists. The blue overlay illustrates the plaque segmentations, while the red overlay shows the scale segmentations. Orange shows the overlap between the two characteristics. While all dermatologists agree in some areas, there are clear disagreements as to which areas contain a certain characteristic.

Table 2: Diagnostic performance (a) and inter-rater agreement (b) on DermX. Dermatologists have high agreement with both the gold standard label and with each other. Seborrhoeic dermatitis stands out as a difficult disease to diagnose, while vitiligo, viral warts and acne appear to be easier.

(a) Dermatologist diagnosis performance with regards to the gold standard (mean \pm std).

	F1	Sensitivity	Specificity
Acne	0.98 \pm 0.01	0.99 \pm 0.01	0.99 \pm 0.01
Actinic keratosis	0.94 \pm 0.05	0.92 \pm 0.08	0.99 \pm 0.01
Psoriasis	0.92 \pm 0.03	0.98 \pm 0.02	0.96 \pm 0.02
Seborrhoeic dermatitis	0.87 \pm 0.07	0.81 \pm 0.11	0.99 \pm 0.01
Viral warts	0.98 \pm 0.02	0.96 \pm 0.03	1.00 \pm 0.00
Vitiligo	0.99 \pm 0.01	0.98 \pm 0.02	1.00 \pm 0.00

(b) Dermatologist inter-rater agreement on diagnosis (mean \pm std).

	F1	Sensitivity	Specificity
Acne	0.95 \pm 0.19	0.95 \pm 0.19	1.00 \pm 0.01
Actinic keratosis	0.90 \pm 0.19	0.91 \pm 0.20	0.99 \pm 0.02
Psoriasis	0.94 \pm 0.07	0.95 \pm 0.09	0.99 \pm 0.02
Seborrhoeic dermatitis	0.90 \pm 0.10	0.92 \pm 0.13	0.99 \pm 0.02
Viral warts	0.99 \pm 0.03	0.99 \pm 0.04	1.00 \pm 0.01
Vitiligo	0.98 \pm 0.03	0.98 \pm 0.05	1.00 \pm 0.01

100 diagnoses, and were disregarded from the analysis. Images where all evaluations were discarded
101 were also removed from the dataset. In the rest of the paper, we will only focus on the remaining
102 1285 evaluations associated with 525 images.

103 The diagnostic accuracy of dermatologists with regards to the gold standard varies between
104 0.92 to 0.99. Seborrhoeic dermatitis is the most difficult disease to diagnose, while vitiligo is
105 the easiest. Pair-wise F1 scores for the inter-rater agreement lies between 0.86 and 1.0 (Table 2).

106 Inter-rater agreement on characteristics (Table 3a) varies significantly more, partially due to the lower
107 number of selections per class. Most basic terms display the highest levels of agreement, with F1
108 scores between 0.65 and 0.88. The two low performing basic terms, macule and nodule, have low
109 selection rates. Several additional terms as defined in Figure 1b, such as open and closed comedones,
110 display levels of agreement similar to the basic terms.

111 Outlining characteristics is a more difficult task, as also confirmed by the low inter-rater F1 scores (Ta-
112 ble 3b). The lower F1 values can also be explained by the annotation protocol specification to
113 prioritise sensitivity over specificity. In terms of sensitivity, we notice the same trend as in the binary
114 agreement: dermatologists tend to agree more on the basic terms. Agreement differences stem from
115 the difficulty in outlining some of these characteristics. For example, comedones cover smaller areas,
116 and dermatologists differed in their approach to outlining them.

Table 3: An inter-rater analysis for supporting characteristics (a) shows significant variation in their 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. Overlap measures (b) show similar differences between raters. Due to the focus on outlining sensitivity at the expense of specificity, most characteristics have a low F1 score. Sensitivity values are high in 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).

(a) Dermatologist inter-rater agreement for the presence or absence of characteristics (mean \pm std).

	F1	Sensitivity	Specificity	Evaluations	Images
Basic terms					
Macule	0.13 \pm 0.24	0.17 \pm 0.31	0.93 \pm 0.10	110	93
Nodule	0.07 \pm 0.22	0.08 \pm 0.26	0.97 \pm 0.05	47	44
Papule	0.65 \pm 0.15	0.69 \pm 0.20	0.86 \pm 0.10	385	213
Patch	0.72 \pm 0.17	0.77 \pm 0.22	0.91 \pm 0.10	335	185
Plaque	0.78 \pm 0.11	0.80 \pm 0.16	0.84 \pm 0.11	592	306
Pustule	0.69 \pm 0.29	0.72 \pm 0.32	0.97 \pm 0.03	161	80
Scale	0.88 \pm 0.09	0.89 \pm 0.12	0.92 \pm 0.09	550	257
Additional terms					
Closed comedo	0.52 \pm 0.27	0.61 \pm 0.35	0.96 \pm 0.05	108	63
Cyst	0.06 \pm 0.22	0.06 \pm 0.23	0.99 \pm 0.02	16	14
Leukotrichia	0.18 \pm 0.38	0.18 \pm 0.38	1.00 \pm 0.01	12	8
Open comedo	0.65 \pm 0.30	0.71 \pm 0.34	0.97 \pm 0.05	132	73
Scar	0.45 \pm 0.29	0.54 \pm 0.38	0.95 \pm 0.06	112	74
Sun damage	0.46 \pm 0.39	0.49 \pm 0.43	0.97 \pm 0.04	101	63
Telangiectasia	0.08 \pm 0.25	0.09 \pm 0.27	0.99 \pm 0.02	13	10
Thrombosed capillaries	0.31 \pm 0.40	0.35 \pm 0.45	0.97 \pm 0.05	67	38

(b) Dermatologist inter-rater localisation agreement for localisable characteristics (mean \pm std).

	F1	Sensitivity	Specificity
Basic terms			
Macule	0.04 \pm 0.12	0.08 \pm 0.20	0.95 \pm 0.13
Nodule	0.03 \pm 0.15	0.06 \pm 0.22	0.98 \pm 0.04
Papule	0.20 \pm 0.28	0.33 \pm 0.36	0.96 \pm 0.10
Patch	0.45 \pm 0.40	0.59 \pm 0.39	0.93 \pm 0.12
Plaque	0.48 \pm 0.39	0.62 \pm 0.37	0.93 \pm 0.12
Pustule	0.24 \pm 0.23	0.38 \pm 0.33	0.99 \pm 0.03
Scale	0.48 \pm 0.32	0.60 \pm 0.33	0.94 \pm 0.10
Additional terms			
Closed comedo	0.08 \pm 0.17	0.24 \pm 0.36	0.93 \pm 0.15
Cyst	0.04 \pm 0.13	0.08 \pm 0.18	1.00 \pm 0.01
Dermatoglyph-disruption	0.33 \pm 0.41	0.48 \pm 0.42	0.98 \pm 0.04
Leukotrichia	0.31 \pm 0.33	0.45 \pm 0.38	0.96 \pm 0.06
Open comedo	0.14 \pm 0.19	0.29 \pm 0.33	0.93 \pm 0.15
Scar	0.12 \pm 0.23	0.26 \pm 0.36	0.91 \pm 0.14
Sun damage	0.35 \pm 0.43	0.51 \pm 0.45	0.75 \pm 0.28
Telangiectasia	0.06 \pm 0.16	0.13 \pm 0.25	0.97 \pm 0.05
Thrombosed capillaries	0.21 \pm 0.30	0.36 \pm 0.38	0.99 \pm 0.02

117 Overall, the contrast between high agreement on diagnoses and low agreement on supporting char-
 118 acteristics illustrates how different experts perceive explanations in different ways. Although they
 119 generally agree on the diagnosis, dermatologists focus on different characteristics to explain their

Table 4: Data used for training and testing the methods, split by disease. An internal clinical dataset was employed for training the models, while DermX was used for testing.

	Acne	Actinic keratosis	Psoriasis	Seborrhoeic dermatitis	Viral warts	Vitiligo
Training	1177	165	975	113	606	178
DermX	99	91	97	78	74	86

120 decision. To properly evaluate a model’s explanations, we must therefore consider the opinions of
 121 multiple experts.

122 3 Explainability Benchmark for Two Architectures

123 Using the DermX dataset, we evaluate the explainability of ConvNets trained for skin lesion diagnosis
 124 by applying Grad-CAM on two models, ResNet-50 and EfficientNet-B4, and comparing the results to
 125 the DermX explanation maps. We selected Grad-CAM for generating the models’ attention maps due
 126 to its high prevalence in the medical image analysis literature [8, 23]. The Keras [24] implementation
 127 of these experiments is available at <https://github.com/leoילב/dermx-experiments>.

128 3.1 Experimental Setup

129 Both architectures were pre-trained on ImageNet and fine-tuned on 3214 images of the six target skin
 130 conditions from an internal clinical dataset. Images from this dataset and their associated diagnoses
 131 were obtained during face-to-face consultations with a dermatologist. All patients included in this
 132 dataset gave their consent for both research and commercial use of their images. Each model was
 133 trained five times, and the results presented are the mean over all trained models.

134 A hyper-parameter search was run on an 80/20 training/validation split for the internal dataset. We
 135 investigated data augmentation parameters (rotation, shear, zoom, brightness), learning rates, and
 136 the number of layers to be fine-tuned. Once the optimal hyper-parameter setup was found (Appendix
 137 Table 2), the two architectures were trained on the entire internal dataset defined in Table 4.
 138 The validation F1 score on the internal dataset was 0.73 for ResNet-50 and 0.79 for EfficientNet-B4.
 139 Finally, the models were tested on the 525 DermX images. All experiments were performed on AWS
 140 EU (Ireland) instances, summing up to two GPU weeks (NVIDIA V100).

141 3.2 Results

142 The expected impact of the data distribution shift was made obvious by the model diagnostic
 143 performance. Diagnostic accuracy with respect to the gold standard is 0.34 ± 0.03 , and 0.42 ± 0.09
 144 for ResNet-50 and EfficientNet-B4, respectively (Table 5). Vitiligo is predicted with both the lowest
 145 sensitivity as well as F1 score for both models, while the highest-ranked diagnosis class for both
 146 models was acne. As seen in Table 5, EfficientNet-B4 outperformed the ResNet-50 on four out of six
 147 diseases, with a difference of 13.5 points in average for F1 score.

148 We evaluate the explainability of the two ConvNets by comparing their attention maps to the
 149 characteristic segmentations. The union of all characteristics segmented by a dermatologist for
 150 an image was also compared to the attention map, as a way to check whether the models take into
 151 account the entire area selected by dermatologists as important to their decision. To quantify the
 152 similarity between the attention maps and the expert-generated maps, we compute the F1 score,
 153 sensitivity and specificity following their fuzzy implementation defined in Crum et al. [25] (Appendix
 154 Table 3).

155 Similar to the diagnostic performance, the explainability of EfficientNet-B4 is higher than that of
 156 ResNet-50 in terms of both F1 score and sensitivity. However, ResNet-50 outperforms EfficientNet-
 157 B4 in terms of specificity on most characteristics and on the union of all characteristics (Table 6).
 158 These observations are also apparent upon visual inspection of the dermatologists segmentations
 159 created by dermatologists and the Grad-CAM visualisations in Figure 3. Much like dermatologists,
 160 both models have higher sensitivity scores for basic terms, albeit at a smaller difference. Within

Table 5: Model diagnostic performance with regards to the gold standard, aggregated over five models. ResNet-50 (a) is out-performed on four out of six diseases by EfficientNet-B4 (b). The training data impact can be seen in the high scores for acne and low scores for vitiligo for both models.

(a) ResNet-50 diagnostic performance with regards to the gold standard (mean \pm std).

	F1	Sensitivity	Specificity
Acne	0.53 ± 0.11	0.43 ± 0.14	0.96 ± 0.02
Actinic keratosis	0.32 ± 0.11	0.23 ± 0.12	0.97 ± 0.02
Psoriasis	0.44 ± 0.04	0.78 ± 0.20	0.60 ± 0.18
Seborrhoeic dermatitis	0.39 ± 0.19	0.41 ± 0.28	0.92 ± 0.08
Viral warts	0.15 ± 0.06	0.14 ± 0.07	0.86 ± 0.04
Vitiligo	0.04 ± 0.02	0.03 ± 0.02	0.90 ± 0.05

(b) EfficientNet-B4 diagnostic performance with regards to the gold standard (mean \pm std).

	F1	Sensitivity	Specificity
Acne	0.65 ± 0.32	0.62 ± 0.33	0.97 ± 0.02
Actinic keratosis	0.55 ± 0.11	0.45 ± 0.15	0.96 ± 0.02
Psoriasis	0.57 ± 0.12	0.90 ± 0.09	0.67 ± 0.23
Seborrhoeic dermatitis	0.45 ± 0.22	0.41 ± 0.22	0.95 ± 0.03
Viral warts	0.07 ± 0.04	0.06 ± 0.04	0.86 ± 0.04
Vitiligo	0.00 ± 0.01	0.00 ± 0.01	0.90 ± 0.03

161 additional terms, cyst, scar, and sun damage all reach sensitivity levels similar to basic terms. This
162 may be due to lower selection rates, as is the case for cyst, or because of the larger areas covered by
163 scar and sun damage in images.

164 4 Discussion and Conclusion

165 Our experiments showcase the intended use of DermX: as an explainability benchmark for der-
166 matological diagnosis ConvNets. By comparing the model explanations to those of the experts
167 not only can we identify the most promising research directions, but also learn about strategies
168 to improve the existing models. For example, if models under consideration systematically miss
169 certain characteristics (i.e. express near-zero sensitivity), one solution is to ensure that training data
170 represents the characteristic well enough by including both positive and negative samples. Another
171 possible outcome is that models consistently highlight different areas than humans (i.e. express low
172 specificity). In this case, ensuring the models are not learning irrelevant characteristics might be done
173 by appropriate training data augmentation. Alternatively, if domain experts confirm that the areas
174 highlighted are relevant for the diagnosis, this knowledge might be used to better educate humans,
175 similar to the actinic keratosis seminar held by Tschandl et al. [8].

176 Our benchmarking results demonstrate that there is still a considerable gap among explanations
177 provided by the models trained for this task and the expert dermatologists. For example, the highest
178 sensitivity achieved for a characteristic by a model on the benchmark is 0.44 ± 0.03 for scale by
179 EfficientNet-B4, which is still significantly below the expert agreement of 0.60 ± 0.33 . Building
180 models that can reach expert level, both in terms of the diagnostic performance and the diagnostic
181 reasoning, would require incorporating such expert annotations in the training process. One solution
182 could be using characteristics maps to guide the attention of a model towards the clinically relevant
183 areas in an image. However, collection of such data is a challenging and laborious task, requiring
184 multiple highly trained dermatologists to meticulously segment and tag the data with a rich set of
185 characteristics.

186 The first release of DermX presented in this work has several limitations. First, only a small number
187 of conditions was selected, which, although highly prevalent [20], are not representative of the whole
188 variety of dermatological diseases. One risk associated with this selection is that future explainable
189 models may focus on this smaller set, at the expense of other, more dangerous conditions. Second,
190 expert annotations were limited to up to three dermatological evaluations per image. Diagnostic

Table 6: Explainability of ResNet-50 (a) and EfficientNet-B4 (b) as similarity measures between dermatologists-segmented supporting characteristics and model activation maps. For each image, the union of all dermatologist characteristic maps was also compared against the activation maps. All activation maps were computed with regards to the gold standard diagnosis using Grad-CAM.

(a) Explainability of ResNet-50 model (mean \pm std).

	F1	Sensitivity	Specificity
Basic terms			
Macule	0.07 \pm 0.02	0.15 \pm 0.02	0.88 \pm 0.02
Nodule	0.05 \pm 0.01	0.19 \pm 0.04	0.88 \pm 0.02
Papule	0.06 \pm 0.01	0.17 \pm 0.02	0.88 \pm 0.02
Patch	0.13 \pm 0.01	0.12 \pm 0.01	0.89 \pm 0.03
Plaque	0.18 \pm 0.05	0.19 \pm 0.04	0.89 \pm 0.01
Pustule	0.02 \pm 0.01	0.21 \pm 0.08	0.87 \pm 0.02
Scale	0.16 \pm 0.05	0.21 \pm 0.05	0.88 \pm 0.01
Additional terms			
Closed comedo	0.07 \pm 0.01	0.15 \pm 0.03	0.87 \pm 0.03
Cyst	0.03 \pm 0.01	0.21 \pm 0.05	0.86 \pm 0.03
Dermatoglyph disruption	0.06 \pm 0.06	0.09 \pm 0.12	0.90 \pm 0.03
Leukotrichia	0.11 \pm 0.02	0.14 \pm 0.03	0.90 \pm 0.01
Open comedo	0.08 \pm 0.01	0.15 \pm 0.03	0.87 \pm 0.03
Scar	0.13 \pm 0.04	0.16 \pm 0.04	0.88 \pm 0.02
Sun damage	0.22 \pm 0.04	0.14 \pm 0.03	0.92 \pm 0.06
Telangiectasia	0.10 \pm 0.02	0.16 \pm 0.06	0.90 \pm 0.04
Thrombosed capillaries	0.03 \pm 0.03	0.10 \pm 0.16	0.91 \pm 0.03
Union	0.17 \pm 0.01	0.16 \pm 0.01	0.90 \pm 0.00

(b) Explainability of EfficientNet-B4 model (mean \pm std).

	F1	Sensitivity	Specificity
Basic terms			
Macule	0.09 \pm 0.03	0.27 \pm 0.07	0.80 \pm 0.03
Nodule	0.03 \pm 0.01	0.20 \pm 0.08	0.82 \pm 0.03
Papule	0.07 \pm 0.01	0.23 \pm 0.07	0.80 \pm 0.03
Patch	0.21 \pm 0.04	0.28 \pm 0.06	0.80 \pm 0.03
Plaque	0.29 \pm 0.01	0.38 \pm 0.03	0.81 \pm 0.04
Pustule	0.02 \pm 0.01	0.28 \pm 0.14	0.82 \pm 0.05
Scale	0.26 \pm 0.01	0.44 \pm 0.03	0.80 \pm 0.04
Additional terms			
Closed comedo	0.08 \pm 0.03	0.21 \pm 0.10	0.83 \pm 0.04
Cyst	0.02 \pm 0.01	0.20 \pm 0.14	0.85 \pm 0.03
Dermatoglyph disruption	0.05 \pm 0.02	0.09 \pm 0.05	0.79 \pm 0.06
Leukotrichia	0.07 \pm 0.03	0.14 \pm 0.10	0.82 \pm 0.04
Open comedo	0.08 \pm 0.04	0.19 \pm 0.10	0.83 \pm 0.04
Scar	0.16 \pm 0.09	0.25 \pm 0.13	0.82 \pm 0.03
Sun damage	0.42 \pm 0.05	0.31 \pm 0.05	0.90 \pm 0.02
Telangiectasia	0.14 \pm 0.01	0.35 \pm 0.04	0.79 \pm 0.04
Thrombosed capillaries	0.01 \pm 0.01	0.07 \pm 0.02	0.80 \pm 0.05
Union	0.25 \pm 0.02	0.29 \pm 0.04	0.82 \pm 0.03

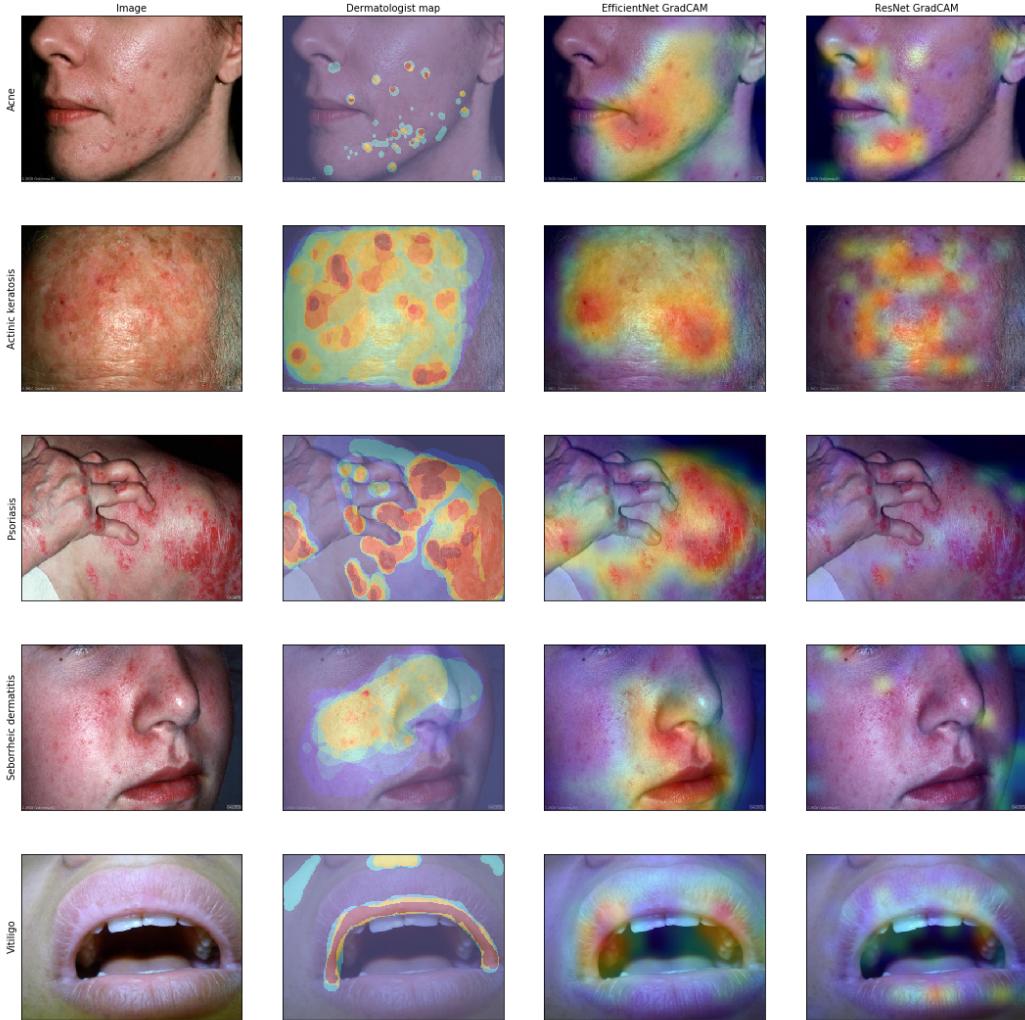


Figure 3: Examples of explanation for images where both models correctly predicted the gold standard diagnosis. From left to right: the original image, the union of all characteristics selected by all dermatologists labelling the image, an EfficientNet-B4 Grad-CAM visualisation, and a ResNet-50 Grad-CAM visualisation. In all cases, the EfficientNet-B4 visualisation is closer to the dermatologist map than the ResNet-50 visualisation. ResNet-50 appears to be more specific, focusing on smaller, more noticeable lesions. More examples can be found in Appendix Figures 4, 5, and 6.

191 reasoning is not a simple task, and is subject to inter-rater variability as seen in our analysis in
 192 Section 2.4. Increasing the number of the dermatologists per image will help make the measurements
 193 more objective. Finally, in terms of the characteristics chosen, the labelling dermatologists could not
 194 select the absence of a characteristic as an important factor in their diagnosis decision.

195 In the future, we aim to continuously expand the dataset with more datapoints to enable training of
 196 diagnostic models along with learning the supportive characteristics. The dataset will also be enriched
 197 with more conditions and more dermatologists to make the next DermX releases more comprehensive
 198 and objective. We will also expand our labelling protocol by including characteristic negation, and
 199 thus expanding the explainability from only supporting characteristics to counterfactual reasoning.

200 To conclude, we introduce DermX, the first dermatological dataset created for diagnosis explainability.
 201 We expect it to serve as a benchmark to meaningfully improve the performance of the ConvNets built
 202 for dermatological diagnosis, and as a possible basis for explainable diagnosis models.

203 **References**

- 204 [1] Christopher J Kelly, Alan Karthikesalingam, Mustafa Suleyman, Greg Corrado, and Dominic
205 King. Key challenges for delivering clinical impact with artificial intelligence. *BMC Medicine*,
206 17(1):1–9, 2019.
- 207 [2] Daniel G Federman, John Concato, and Robert S Kirsner. Comparison of dermatologic
208 diagnoses by primary care practitioners and dermatologists: a review of the literature. *Archives
209 of Family Medicine*, 8(2):170, 1999.
- 210 [3] Hao Feng, Juliana Berk-Krauss, Paula W Feng, and Jennifer A Stein. Comparison of dermatol-
211 ogist density between urban and rural counties in the United States. *JAMA Dermatology*, 154
212 (11):1265–1271, 2018.
- 213 [4] Dionne S Kringos, Wienke GW Boerma, Allen Hutchinson, Richard B Saltman, World Health
214 Organization, et al. *Building primary care in a changing Europe*. World Health Organization.
215 Regional Office for Europe, 2015.
- 216 [5] Ayush Jain, David Way, Vishakha Gupta, Yi Gao, Guilherme de Oliveira Marinho, Jay Hartford,
217 Rory Sayres, Kimberly Kanada, Clara Eng, Kunal Nagpal, et al. Development and assessment of
218 an artificial intelligence-based tool for skin condition diagnosis by primary care physicians and
219 nurse practitioners in teledermatology practices. *JAMA Network Open*, 4(4):e217249–e217249,
220 2021.
- 221 [6] Kenneth Thomsen, Lars Iversen, Therese Louise Titlestad, and Ole Winther. Systematic review
222 of machine learning for diagnosis and prognosis in dermatology. *Journal of Dermatological
223 Treatment*, 31(5):496–510, 2020.
- 224 [7] Andre Esteva, Brett Kuprel, Roberto A Novoa, Justin Ko, Susan M Swetter, Helen M Blau, and
225 Sebastian Thrun. Dermatologist-level classification of skin cancer with deep neural networks.
226 *Nature*, 542(7639):115–118, 2017.
- 227 [8] Philipp Tschandl, Christoph Rinner, Zoe Apalla, Giuseppe Argenziano, Noel Codella, Allan
228 Halpern, Monika Janda, Aimilios Lallas, Caterina Longo, Josep Malvehy, et al. Human-
229 computer collaboration for skin cancer recognition. *Nature Medicine*, 26(8):1229–1234, 2020.
- 230 [9] Bryce Goodman and Seth Flaxman. European union regulations on algorithmic decision-making
231 and a “right to explanation”. *AI magazine*, 38(3):50–57, 2017.
- 232 [10] Eric J Topol. High-performance medicine: the convergence of human and artificial intelligence.
233 *25(1):44–56*, 2019.
- 234 [11] Marco Tulio Ribeiro, Sameer Singh, and Carlos Guestrin. "Why should i trust you?" Explaining
235 the predictions of any classifier. In *Proceedings of the 22nd ACM SIGKDD international
236 conference on knowledge discovery and data mining*, pages 1135–1144, 2016.
- 237 [12] Scott M Lundberg and Su-In Lee. A unified approach to interpreting model predictions. In
238 I. Guyon, U. V. Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan, and R. Garnett,
239 editors, *Advances in Neural Information Processing Systems*, volume 30. Curran Associates,
240 Inc., 2017.
- 241 [13] Ramprasaath R Selvaraju, Michael Cogswell, Abhishek Das, Ramakrishna Vedantam, Devi
242 Parikh, and Dhruv Batra. Grad-CAM: Visual explanations from deep networks via gradient-
243 based localization. In *Proceedings of the IEEE International Conference on Computer Vision*,
244 pages 618–626, 2017.
- 245 [14] Philipp Tschandl, Cliff Rosendahl, and Harald Kittler. The HAM10000 dataset, a large collection
246 of multi-source dermatoscopic images of common pigmented skin lesions. *Scientific Data*, 5(1):
247 1–9, 2018.
- 248 [15] Alexander Nast, Chris E M Griffiths, Roderick Hay, Wolfram Sterry, and Jean L Bolognia. The
249 2016 International League of Dermatological Societies' revised glossary for the description of
250 cutaneous lesions. *British Journal of Dermatology*, 174(6):1351–1358, 2016.

- 251 [16] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for im-
252 age recognition. In *Proceedings of the IEEE Conference on Computer Vision and Pattern*
253 *Recognition*, pages 770–778, 2016.
- 254 [17] Mingxing Tan and Quoc Le. EfficientNet: Rethinking model scaling for convolutional neural
255 networks. In *International Conference on Machine Learning*, pages 6105–6114. PMLR, 2019.
- 256 [18] Dermnetnz. <https://dermnetnz.org/>, 2021. Accessed: 2021-04-01.
- 257 [19] Xiaoxiao Sun, Jufeng Yang, Ming Sun, and Kai Wang. A benchmark for automatic visual
258 classification of clinical skin disease images. In *European Conference on Computer Vision*,
259 pages 206–222. Springer, 2016.
- 260 [20] Henry W Lim, Scott AB Collins, Jack S Resneck Jr, Jean L Bolognia, Julie A Hodge, Thomas A
261 Rohrer, Marta J Van Beek, David J Margolis, Arthur J Sober, Martin A Weinstock, et al. The
262 burden of skin disease in the united states. *Journal of the American Academy of Dermatology*,
263 76(5):958–972, 2017.
- 264 [21] Amanda Oakley. *Dermatology Made Easy*. Scion Publishing Ltd, The Old Hayloft, Vantage
265 Business Park, Bloxham Road, Banbury OX16 9UX, UK, 2017.
- 266 [22] Darwin v7 labs. <https://darwin.v7labs.com>, 2021. Accessed: 2021-05-01.
- 267 [23] Jeremy Irvin, Pranav Rajpurkar, Michael Ko, Yifan Yu, Silviana Ciurea-Ilcus, Chris Chute,
268 Henrik Marklund, Behzad Haghgoo, Robyn Ball, Katie Shpanskaya, et al. CheXpert: A large
269 chest radiograph dataset with uncertainty labels and expert comparison. In *Proceedings of the*
270 *AAAI Conference on Artificial Intelligence*, volume 33, pages 590–597, 2019.
- 271 [24] François Chollet et al. Keras. <https://keras.io>, 2015.
- 272 [25] William R Crum, Oscar Camara, and Derek L G Hill. Generalized overlap measures for
273 evaluation and validation in medical image analysis. *IEEE Transactions on Medical Imaging*,
274 25(11):1451–1461, 2006.

275 **Checklist**

- 276 1. For all authors...
 - 277 (a) Do the main claims made in the abstract and introduction accurately reflect the paper's
278 contributions and scope? **[Yes]**
 - 279 (b) Did you describe the limitations of your work? **[Yes]** See Section 4.
 - 280 (c) Did you discuss any potential negative societal impacts of your work? **[Yes]** See
281 Section 4.
 - 282 (d) Have you read the ethics review guidelines and ensured that your paper conforms to
283 them? **[Yes]**
- 284 2. If you are including theoretical results...
 - 285 (a) Did you state the full set of assumptions of all theoretical results? **[N/A]**
 - 286 (b) Did you include complete proofs of all theoretical results? **[N/A]**
- 287 3. If you ran experiments...
 - 288 (a) Did you include the code, data, and instructions needed to reproduce the main experi-
289 mental results (either in the supplemental material or as a URL)? **[Yes]** See Section 3
290 and Appendix Table 2.
 - 291 (b) Did you specify all the training details (e.g., data splits, hyperparameters, how they
292 were chosen)? **[Yes]** See Section 3.1.
 - 293 (c) Did you report error bars (e.g., with respect to the random seed after running experi-
294 ments multiple times)? **[Yes]** All results described in Section 3.2 are computed over
295 five training runs, and are reported as mean \pm standard deviation.
 - 296 (d) Did you include the total amount of compute and the type of resources used (e.g., type
297 of GPUs, internal cluster, or cloud provider)? **[Yes]** See Section 3.1.
- 298 4. If you are using existing assets (e.g., code, data, models) or curating/releasing new assets...
 - 299 (a) If your work uses existing assets, did you cite the creators? **[Yes]** See Section 2.
 - 300 (b) Did you mention the license of the assets? **[Yes]** See Section 2.
 - 301 (c) Did you include any new assets either in the supplemental material or as a URL? **[Yes]**
302 See Abstract 2.
 - 303 (d) Did you discuss whether and how consent was obtained from people whose data you're
304 using/curating? **[Yes]** See Section 2.
 - 305 (e) Did you discuss whether the data you are using/curating contains personally identifiable
306 information or offensive content? **[Yes]** In Section 2, we discuss the presence of
307 personally identifiable data in the dataset images. As all diseases included in the dataset
308 often manifest on the face, it was intractable to exclude such images from the set.
- 309 5. If you used crowdsourcing or conducted research with human subjects...
 - 310 (a) Did you include the full text of instructions given to participants and screenshots, if
311 applicable? **[Yes]** See Section 2.3.
 - 312 (b) Did you describe any potential participant risks, with links to Institutional Review
313 Board (IRB) approvals, if applicable? **[No]** The participant data we use has already
314 been made public by DermNetNZ and SD-260. Annotation data does not expose any
315 information about the labellers involved in the project, and thus an IRB approval was
316 not deemed applicable.
 - 317 (c) Did you include the estimated hourly wage paid to participants and the total amount
318 spent on participant compensation? **[No]** All eight dermatologists that helped label this
319 dataset are hired as consultants by Omhu. Their salaries are confidential information.

320 **A Appendix**

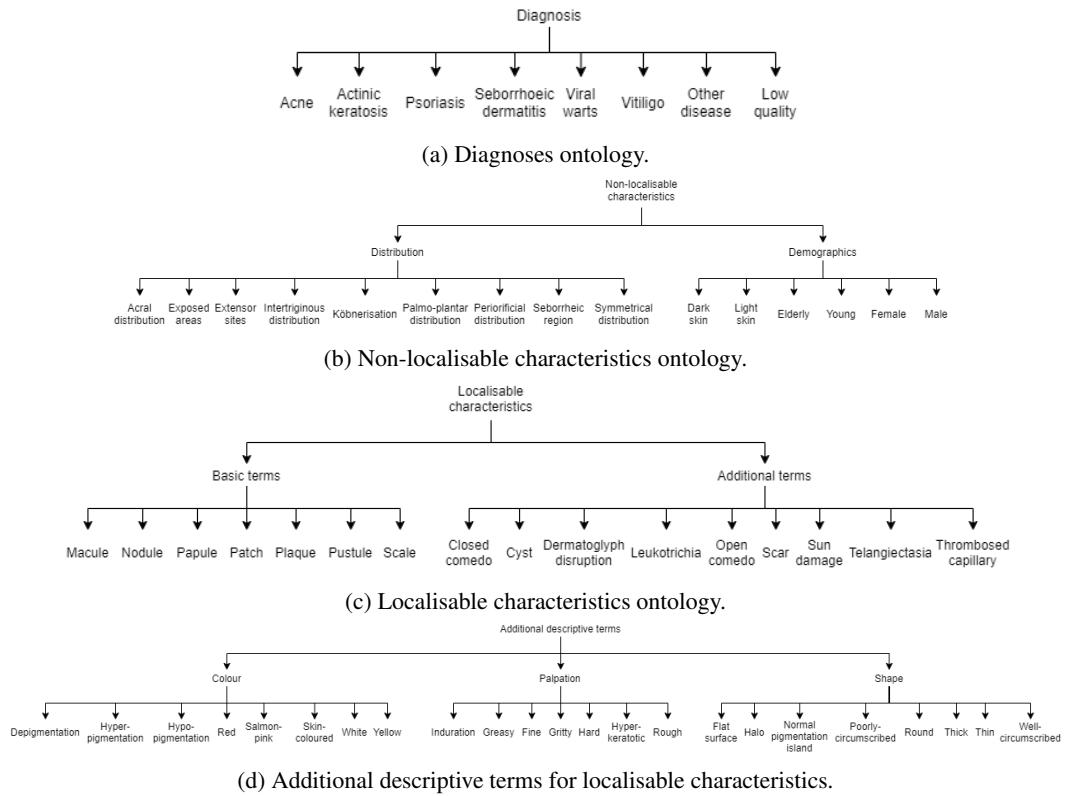


Figure 1: Ontology of the four types of labels. The list of diagnoses (1a) includes the six diseases and two discard options for images that either displayed another disease or were of low quality. Non-localisable characteristics (1b) were added to the ILDS classification as global image tags after being flagged as relevant by our senior dermatologists. Localisable characteristics (1c) and additional descriptive terms (1d) were tailored for the six diseases from medical resources [15, 21], and with the help of two senior dermatologists.

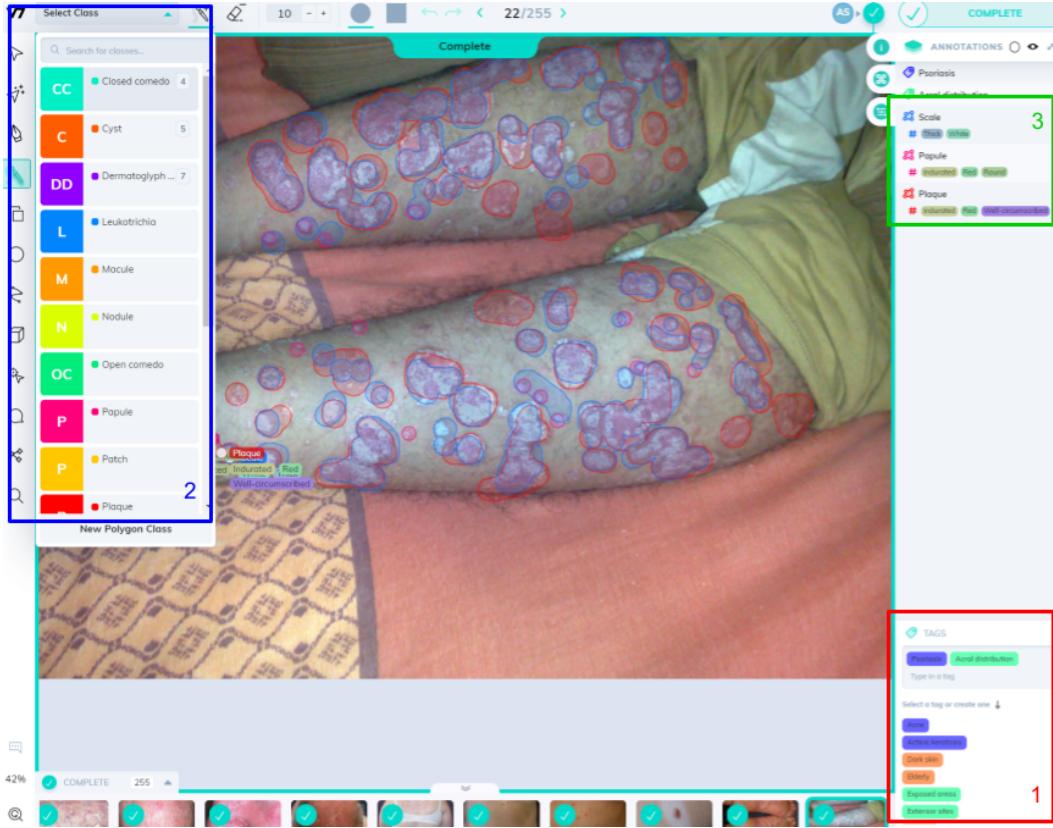
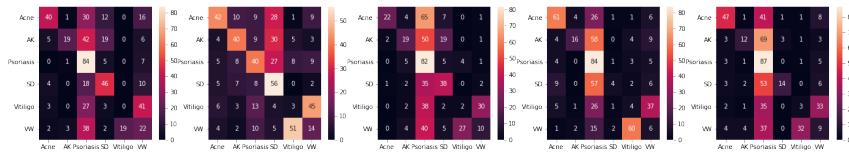
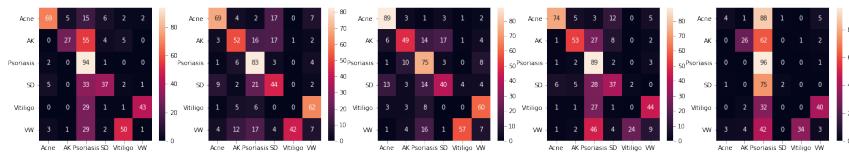


Figure 2: Labelling tool interface, exemplified for a psoriasis case from the SD-260 dataset. In the global tag search box (area 1, bottom right), dermatologists can select the disease, relevant demographics information, and lesion distribution. The brush selection menu (area 2, top left) allows them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



(a) Confusion matrix for each ResNet-50 runs.



(b) Confusion matrix for each EfficientNet-B4 runs.

Figure 3: Confusion matrix for all models trained. Both ResNet-50 (a) and EfficientNet-B4 (b) show a bias towards predicting psoriasis, and predict vitiligo in very few cases.

Table 1: Dermatologist inter-rater agreement for the presence or absence of characteristics, including the number of evaluations (evals) and the number of images where they were identified.

	F1	Sensitivity	Specificity	Evals	Images
Non-localisable characteristics					
Demographics					
Elderly	0.50 ± 0.29	0.58 ± 0.37	0.93 ± 0.09	186	117
Young	0.29 ± 0.29	0.36 ± 0.39	0.9 ± 0.13	168	123
Female	0.14 ± 0.23	0.18 ± 0.33	0.94 ± 0.10	84	67
Male	0.14 ± 0.22	0.20 ± 0.35	0.91 ± 0.13	140	113
Dark skin	0.00 ± 0.00	0.00 ± 0.00	0.97 ± 0.05	30	30
Light skin	0.10 ± 0.22	0.14 ± 0.31	0.82 ± 0.27	222	193
Distribution					
Acral distribution	0.33 ± 0.29	0.38 ± 0.38	0.92 ± 0.08	149	100
Exposed areas	0.47 ± 0.33	0.54 ± 0.38	0.89 ± 0.12	255	172
Extensor sites	0.28 ± 0.32	0.31 ± 0.38	0.95 ± 0.06	85	59
Intertriginous	0.00 ± 0.00	0.00 ± 0.00	0.99 ± 0.01	9	9
Köbnerization	0.05 ± 0.20	0.06 ± 0.23	0.98 ± 0.02	19	17
Palmo-plantar	0.31 ± 0.33	0.36 ± 0.41	0.97 ± 0.04	74	52
Periorificial	0.16 ± 0.35	0.16 ± 0.36	0.99 ± 0.02	24	18
Seborrhoeic region	0.66 ± 0.24	0.74 ± 0.30	0.92 ± 0.1	287	160
Symmetrical	0.21 ± 0.24	0.26 ± 0.33	0.93 ± 0.07	105	85
Localisable characteristics					
Basic terms					
Macule	0.13 ± 0.24	0.17 ± 0.31	0.93 ± 0.1	110	93
Nodule	0.07 ± 0.22	0.08 ± 0.26	0.97 ± 0.05	47	44
Papule	0.65 ± 0.15	0.69 ± 0.20	0.86 ± 0.10	385	213
Patch	0.72 ± 0.17	0.77 ± 0.22	0.91 ± 0.10	335	185
Plaque	0.78 ± 0.11	0.80 ± 0.16	0.84 ± 0.11	592	306
Pustule	0.69 ± 0.29	0.72 ± 0.32	0.97 ± 0.03	161	80
Scale	0.88 ± 0.09	0.89 ± 0.12	0.92 ± 0.09	550	257
Additional terms					
Closed comedo	0.52 ± 0.27	0.61 ± 0.35	0.96 ± 0.05	108	63
Cyst	0.06 ± 0.22	0.06 ± 0.23	0.99 ± 0.02	16	14
Dermatoglyph disruption	0.32 ± 0.37	0.33 ± 0.39	0.97 ± 0.04	86	50
Leukotrichia	0.18 ± 0.38	0.18 ± 0.38	1.00 ± 0.01	12	8
Open comedo	0.65 ± 0.30	0.71 ± 0.34	0.97 ± 0.05	132	73
Scar	0.45 ± 0.29	0.54 ± 0.38	0.95 ± 0.06	112	74
Sun damage	0.46 ± 0.39	0.49 ± 0.43	0.97 ± 0.04	101	63
Telangiectasia	0.08 ± 0.25	0.09 ± 0.27	0.99 ± 0.02	13	10
Thrombosed capillary	0.31 ± 0.40	0.35 ± 0.45	0.97 ± 0.05	67	38

Table 2: Optimal hyper-parameter setup and other training parameters for ResNet-50 and EfficientNet-B4, as identified after a hyper-parameter search.

	ResNet-50	EfficientNet-B4
Rotation	20	20
Shear	0	0.5
Zoom	0.25	0.5
Brightness	0.25-1	0.5-1
Learning rate	0.01	0.001
Optimiser	Adam	Adam
Training epochs	30	15
Image size	300 × 400	300 × 400
Weighted classes	On	On

Table 3: Similarity metrics used for comparison of models attention maps (\mathcal{A}) and dermatologists characteristics segmentations (\mathcal{S}).

Similarity metric	Formula
F1 score	$\frac{2 \sum_{p \in pixels} \min(\mathcal{A}_p, \mathcal{S}_p)}{\sum_{p \in pixels} (\mathcal{A}_p) + \sum_{p \in pixels} (\mathcal{S}_p)}$
Sensitivity	$\frac{\sum_{p \in pixels} \min(\mathcal{A}_p, \mathcal{S}_p)}{\sum_{p \in pixels} (\mathcal{S}_p)}$
Specificity	$\frac{\sum_{p \in pixels} \min(1 - \mathcal{A}_p, 1 - \mathcal{S}_p)}{\sum_{p \in pixels} (1 - \mathcal{S}_p)}$

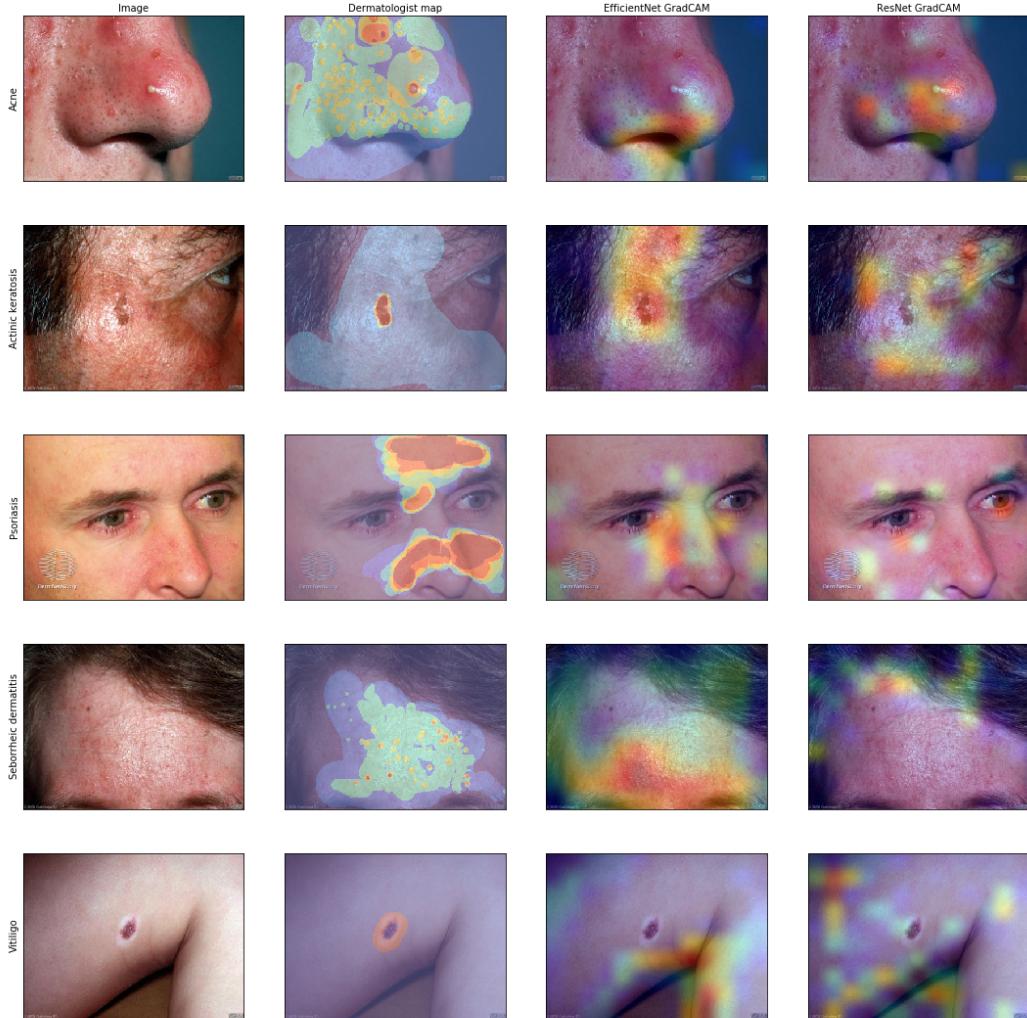


Figure 4: Explanation for images where ResNet correctly predicted the class, while EfficientNet did not. From left to right: the original image, the union of all characteristics selected by all dermatologists labelling the image, an EfficientNetB4 Grad-CAM visualisation, and a ResNet-50 Grad-CAM visualisation.

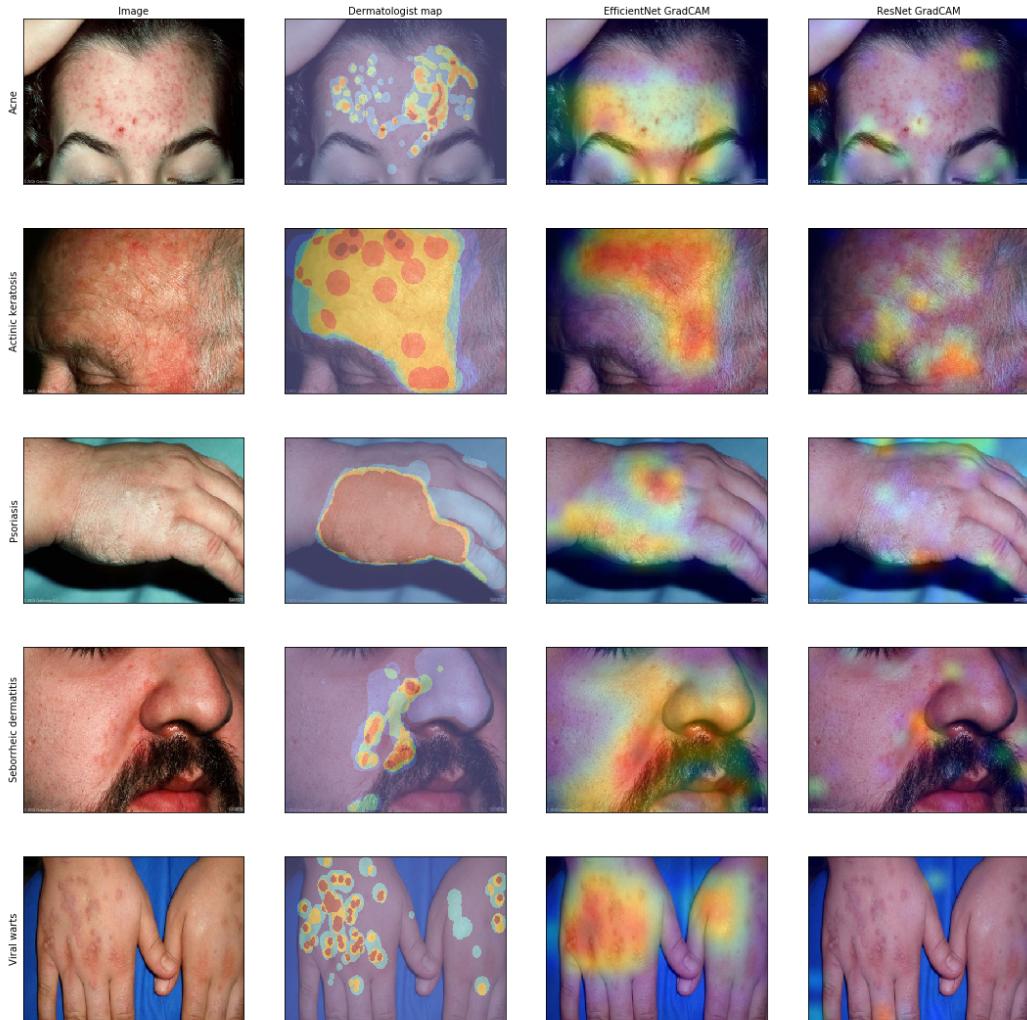


Figure 5: Explanation for images where EfficientNet correctly predicted the class, while EfficientNet did not. From left to right: the original image, the union of all characteristics selected by all dermatologists labelling the image, an EfficientNetB4 Grad-CAM visualisation, and a ResNet-50 Grad-CAM visualisation.

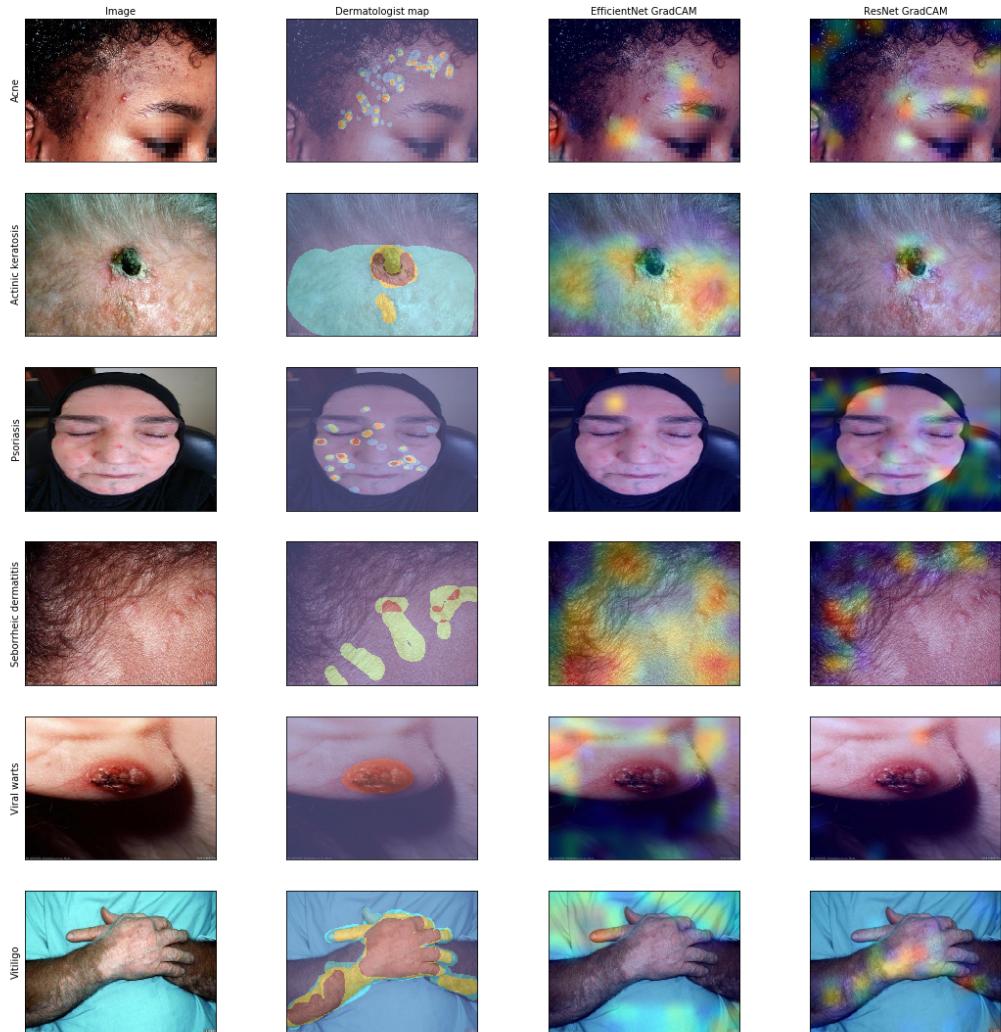


Figure 6: Explanation for images where neither of the two models correctly predicted the class, while EfficientNet did not. From left to right: the original image, the union of all characteristics selected by all dermatologists labelling the image, an EfficientNetB4 Grad-CAM visualisation, and a ResNet-50 Grad-CAM visualisation.