

Automatic SCORing of Atopic Dermatitis using Deep Learning (ASCORAD): A Pilot Study

Short title: Automatic atopic dermatitis severity assessment

Alfonso Medela^{1,*}, Taig Mac Carthy^{2,**,+}, S. Andy Aguilar Robles^{1,***},

Carlos M Chiesa-Estomba^{3,4,5}, and Ramon Grimalt^{6,+}

¹Department of Medical Computer Vision and PROMs, Legit.Health, 48013, Bilbao, Spain

²Department of Clinical Endpoint Innovation, Legit.Health, 48013, Bilbao, Spain

³Department of Otorhinolaryngology, Osakidetza, Donostia University Hospital, 20014 San Sebastian, Spain

⁴Biodonostia Health Research Institute, 20014 San Sebastian, Spain

⁵Head Neck Study Group of Young-Otolaryngologists of the International Federations of Otorhino-laryngological Societies (YO-IFOS), 13005 Marseille, France

⁶Facultat de Medicina i Ciències de la Salut. UIC-Barcelona, Universitat Internacional de Catalunya, Sant Cugat del Vallès, Barcelona, Spain

* alfonso@legit.health

** taig@legit.health

*** andy@legit.health

+these authors contributed equally to this work

Corresponding author:

Alfonso Medela

Department of Medical Computer Vision and PROMs

Legit.Health

48013, Bilbao, Spain

alfonso@legit.health

Twitter: @alfonsomedela

Abbreviations:

AD - Atopic Dermatitis; SCORAD - SCORing Atopic Dermatitis; ASCORAD - Automatic SCORing Atopic Dermatitis; CADx - Computer-Aided Diagnosis; ETFAD - European Task Force on Atopic Dermatitis; PO - Patient-Oriented; PO-SCORAD - Patient-Oriented SCORing Atopic Dermatitis; TIS - Three Item Severity; EASI - Eczema Area and Severity Index; MPT - Multiphoton Tomography; EHR - Electronic Health Record; PASI - Psoriasis Area and Severity Index; CNN - Convolutional Neural Network; MP - Megapixel; JPEG - Joint Photographic Expert Group; DEX - Deep EXpectation; RSD - Relative Standard Deviation; FAR - Full Agreement Rate; PAR - Partial Agreement Rate; AUC - Area Under the Curve; IoU - Intersection over Union; RMAE - Relative Mean Absolute Error; GPU - Graphics Processing Unit; BSA - Body Surface Area; ACC — Accuracy.

ABSTRACT

Atopic dermatitis (AD) is a chronic, itchy skin condition that affects 15–20% of children, but may occur at any age. It is estimated that 16.5 million U.S. adults (7.3%) have AD that initially began at >2 years of age, with nearly 40% affected by moderate or severe disease. Therefore, a quantitative measurement that could track the evolution of atopic dermatitis severity could be extremely useful in assessing therapeutic efficacy. Currently, SCORAD (SCORing Atopic Dermatitis) is the most frequently used measurement in clinical practice. However, SCORAD has the following disadvantages: (1) Time consuming: calculating SCORAD usually takes about 7–10 minutes per patient which poses a heavy burden on dermatologists; and (2) Inconsistency: due to the complexity of SCORAD calculation, even well-trained dermatologists could give different scores for the same case. In this study we introduce ASCORAD, an automatic version of the SCORAD, based on state-of-the-art convolutional neural networks that measure atopic dermatitis severity based on skin lesion images. Overall, we have demonstrated that ASCORAD may prove to be a rapid and objective alternative method for the automatic assessment of atopic dermatitis, achieving results comparable to human expert assessment while reducing inter-observer variability.

Keywords: Atopic dermatitis, SCORAD, Deep learning, Automatic severity assessment, CADx system

INTRODUCTION

Atopic dermatitis (AD) is a multifaceted, chronic relapsing inflammatory skin disease that is commonly associated with other atopic manifestations such as allergic conjunctivitis, allergic rhinitis, and asthma (Bieber, 2008)(Berke et al., 2012)(Drucker et al., 2017). It is the most common skin disease in children, affecting approximately 15% to 20% of children and 1% to 3% of adults (Eichenfield et al., 2014)(Nuttan, 2015). Onset of disease is most common by 5 years of age, and early diagnosis and treatment are essential to avoid complications of AD and improve quality of life (Eichenfield et al., 2014).

The European Task Force on Atopic Dermatitis (ETFAD) developed the SCORAD (Stalder et al., 1993) (SCORing Atopic Dermatitis) index to create a consensus on assessment methods for AD. The SCORAD index consists of the interpretation of the extent of the disorder, that is, the intensity, composed of six visual items (erythema, oedema/papules, effect of scratching, oozing/crust formation, lichenification, and dryness), and two subjective symptoms (itch and sleeplessness); the maximum score is 103 points. If the subjective symptoms (itch and sleeplessness) of the SCORAD are not assessed; the total score is 83 points and it is known as the objective SCORAD. The SCORAD index is influenced by subjective ratings that may be affected by social and cultural factors and therefore, ETFAD recommends the objective SCORAD. One of the advantages of using the objective SCORAD system is that it is based on a European consensus of experts on pediatric dermatology. The system is representative and well evaluated (Schmitt et al., 2013), but shows, as with all other systems, intra- and inter-observer disagreements. However, it is currently widely used in clinical practise to assess patient evolution and measure the effectiveness of treatments (Panahi et al., 2011)(Butler et al., 2020)(Silverberg et al., 2020)(Yoo et al., 2020)(Nahm et al., 2020).

Much work has been done in the development of a better scoring system to reach a more objective and quicker to fill scoring system. Novel tools for patient use like the patient-oriented (PO) validated scoring system PO-SCORAD (Stalder et al., 2011) detect changes in signs and symptoms without the intervention of doctors. The Three Item Severity score (TIS) is a simple method to determine the severity of AD, and assessment of the total TIS takes about 43 seconds per patient. The eczema area and severity index (EASI) (Hanifin et al., 2001) showed a good inter and intra observer variability but it is a complex and time consuming index to fill. However, all these scoring systems still suffer from the same variability problem as they share similarities with SCORAD (Chopra et al., 2017).

In recent years artificial intelligence has achieved human-expert-like performance in a wide variety of tasks such as skin cancer classification, detection and lesion segmentation. Extensive work has been done in the detection of atopic dermatitis with different imaging methods including MPT (Guimarães et al., 2020), clinical image (Wu et al., 2020), and even electronic health records (EHR) (Gustafson et al., 2017), Skin pathologies like psoriasis have also attracted the attention of researchers for the same reasons as atopic dermatitis, as the main scoring system, PASI, is a time consuming and highly subjective scoring method. Dash et al. (Dash et al., 2019) proved that convolutional neural networks are able to segmentate psoriasis with high accuracy, sensitivity and specificity outperforming existing methods. Anabik Pal et al. (Pal et al., 2016) demonstrated the effectiveness of CNNs in visual sign classification, a key task to automatic severity grading. Dash et al. (Dash et al., 2020) combined both segmentation and severity grading creating a CADx system for psoriasis lesion grading.

Creating a more objective and practical scoring system for atopic dermatitis assessment is key to improve evidence-based dermatology. In this study we introduce ASCORAD, an automatic version of the SCORAD: a quick, accurate and fully automated scoring system.

MATERIALS AND METHODS

Datasets and annotations

In this retrospective, non-interventional study, three new annotated datasets were constructed in order to train and validate the performance of the lesion surface segmentation and visual sign severity assessment algorithms. The first two datasets are of purely light skinned patients (Fitzpatrick I, II y III), as it proved to be easier to gather datasets of such characteristics, whilst the third consists of images of IV, V and VI skin types according to Fitzpatrick scale. Demographic characteristics of each dataset are gathered in Table 1. Clinical images were collected from online public sources and patient consent and ethics committee were not necessary. Published images belong to Danderm dermatology atlas and the author gave his consent for publication.

Legit.Health-AD dataset

Legit.Health-AD is a dataset collected from online dermatological atlases that consists of 604 images that belong to light skinned patients, of which one third are children (Table 1), suffering from atopic dermatitis with lesions present on different body parts. The dataset contains the following percentage of body zones: head (22%), trunk (11%), arms (23%), hands (9%), legs (16%), feet (8%), genitalia (3%), full body (1%) and skin close-up (7%). The dataset contains a substantial variety of clinical images taken from different angles, distances, light conditions, body

parts and disease severity. Figure 1 depicts the normalized intensity distribution by visual sign. The images have a minimum size of 260 x 256, an average size of 667 x 563 and a maximum size of 1772 x 1304.

Legit.Health-AD-Test dataset

A second dataset, *Legit.Health-AD-Test*, was built for testing purposes. The dataset was gathered from several dermatological atlases publicly available and contains a total number of 367 images that belong exclusively to light skinned patients. The dataset is only characterized by skin type (Table 1) and basic demographic information like age and sex is missing as the original sources do not provide that information. The images were downloaded one by one and each of them was reviewed by a physician to approve the inclusion of the image in the dataset. Duplicates or very similar images were removed and no other data sampling technique was applied. Similarly to *Legit.Health-AD*, the dataset contains images of children and adults with a great variability in angles, distances, light conditions, body parts and disease severity. The dataset contains the following percentage of body zones: head (35%), trunk (20%), arms (18%), hands (7%), legs (13%), feet (2%), genitalia (2%) and skin close-up (3%). The visual sign intensity distribution of this dataset is different to *Legit.Health-AD*, having more cases of zero intensity for most of the visual signs (Figure 1). The images have a minimum size of 313 x 210, an average size of 574 x 537 and a maximum size of 2848 x 3252.

Legit.Health-AD-FPK-IVI dataset

Legit.Health-AD-FPK-IVI is a dataset collected from online dermatological atlases that contains photos of children and adult patients with Fitzpatrick IV, V and VI skin types suffering from atopic dermatitis. The same manual procedure as for *Legit.Health-AD-Test* was applied in order to gather

the dataset and basic demographic information like age and sex is also missing (Table 1). It is composed of 112 images with a minimum size of 200 x 204, an average size of 766 x 695 and a maximum size of 3024 x 4032. The dataset contains the following percentage of body zones: head (41%), trunk (10%), arms (17%), hands (8%), legs (13%), feet (3%) and skin close-up (8%). The goal of including this dataset in the study was to gather preliminary evidence of the efficiency of the algorithms in dark skin.

Ground truth labels

The corresponding ground truths of each dataset were prepared by nine experts, three for each dataset, who treat patients with atopic dermatitis in their daily practice, to reduce variability by combining their results. The experts annotated the images without more context than the images. They had to draw a mask over the lesion and choose a score from 0 to 3 for each visual sign that comprise the SCORAD.

We obtained the ground truth labels for lesion segmentation and visual sign intensity classification by averaging the masks of the three annotators and by averaging the intensity levels. We chose the mean over the median as it is the statistical measure that gets the best results for generating ground truth labels from multi-annotator ordinal data (Lakshminarayanan and Teh, 2013).

Data preprocessing

Images were resized to 512×512 and pixel values scaled between zero and one. In addition, images in which the disease was too small in the picture were cropped, focusing on the disease. Ground truth labels were obtained from averaging the results as explained in the previous section. However, we ran some additional experiments using an alternative ground truth only for the

training set, consisting of the median visual intensity, instead of the mean. As a result of applying the mean and median, discrete visual sign intensity levels yielded real numbers, which had to be rounded in order to return to the discrete range [0,3]. To prevent information loss, we considered rescaling the values to [0,10] and [0,100] before rounding, and compared these ranges to the original one.

With regards to lesion surface masks, the average mask was computed, resulting in a grayscale image in the range [0,255]. A pixel intensity threshold of 155 was applied to obtain a binary mask that was used as the ground truth. Images were finally normalised to the range [0,1].

Deep learning model

The ASCORAD calculation can be divided in two parts, lesion surface segmentation and visual sign severity assessment. We trained two separated models, one for each task, and named *Legit.Health-SCORADNet* to the neural networks involved in the calculation of the ASCORAD (source code available at github.com/Legit-Health/ASCORAD).

Lesion surface segmentation

For the lesion surface segmentation problem we applied a U-Net, an architecture that was first designed for biomedical image segmentation and demonstrated great results on the task of cell tracking (Ronneberger et al., 2015). The main contribution of this architecture was the ability to achieve good results even with hundreds of examples. The U-Net consists of two paths: a contracting path and an expanding path. The contracting path is a typical convolutional network where convolution and pooling operations are repeatedly applied. We decided to use the Resnet-34 (He et al., 2015) architecture, which is the typical backbone used in the contracting path.

Visual sign severity assessment

We trained a multi-output (Xu et al., 2020) classifier, with one softmax layer per visual sign (Figure 2). We used the EfficientNet-B0 network architecture (Tan and Lee, 2019) that was pretrained on approximately 1.28 million images (1,000 object categories) from the 2014 ImageNet Large Scale Visual Recognition Challenge (Russakovsky et al., 2014), and trained it on our dataset using transfer learning (Pan et al., 2010). EfficientNets achieve better accuracy and efficiency than previous CNNs with fewer parameters by applying a new scaling method that uniformly scales all dimensions of depth/width/resolution using a simple yet highly effective compound coefficient. There are eight versions consisting of a different number of parameters, with the B0 being the smallest network that achieves state-of-the-art 77.1% top-1 accuracy on ImageNet for a network consisting of 5 million parameters.

Visual sign severity grading can be seen as a piecewise regression, or, alternatively, as a discrete classification with four discrete value labels for each visual sign intensity. In the case of multiple visual signs, a multi-label classification network can be used to solve the problem. However, in order to exploit methods like DEX (Rothe et al., 2015), one softmax layer per visual sign is needed. So, for the purpose of applying the DEX method, we constructed a multi-output classifier with six softmax layers consisting of N neurons each, with N being 4, 11 or 101, depending on the range of the ground truth labels.

DEX method proved to obtain better results on regression metrics by approaching a regression problem like a classification and applying a softmax expected value:

$$E(y) = \sum_{i=0}^3 p_i y_i$$

(1)

where $\mathcal{C} = 0, 1, \dots, C$ is the C -dimensional output layer of each visual sign, representing softmax output probabilities $\mathcal{P}_{\mathcal{C}} \in \mathcal{C}$, and $\mathcal{I}_{\mathcal{C}}$ are the discrete intensity levels corresponding to each class \mathcal{C} .

Evaluation metrics

Dermatologists may have a bias, a fixed effect where one observer consistently measures high or low. There may also be a random effect or heterogeneity, where the observer measures higher than others for some subjects and lower for others. In order to measure inter-observer variability, understand annotation quality in more detail and compare it to the performance of the algorithms, we calculated a set of metrics.

First of all, we computed the Relative Standard Deviation (RSD) and Cohen's kappa for all the visual signs and lesion surface segmentation. In the case of the annotation of visual sign intensity, we also measured the times that the three observers gave the same result or the full agreement rate (FAR). To complement FAR, two more metrics were calculated: the times that at least two observers gave the same result whilst the third observer gave a result that deviated ± 1 from the other observer's, or the Partial Agreement Rate 1 (PAR1). The same metrics without the ± 1 condition for the third observer was called Partial Agreement Rate 2 (PAR2). Therefore, the metrics are ordered as follows in regards to their restrictiveness: $FAR > PAR1 > PAR2$. To assess the quality of the annotations and understand in more depth the results, we compared the results against an algorithm that randomly picked three intensity values for each visual sign. We ran this for millions of times and found that RSD of a random visual sign evaluation tends to 27%, FAR to 6%, PAR1 to 34% and PAR2 to 62%.

We also calculated metrics that allowed a direct comparison of the *Legit.Health-SCORADNet* and annotation, for both lesion segmentation and visual sign severity assessment. Pixel accuracy, Area Under the Curve (AUC), Intersection over Union (IoU) and F1 Score metrics were the preferred metrics for segmentation, whilst for the severity assessment of visual signs we used Relative Mean Absolute Error (RMAE).

Experimental setup

We ran two main experiments for each task, one with images containing only light skin and another adding a small number of dark skin images in the training set.

In the first experiment, we used *Legit.Health-AD* for training, and *Legit.Health-AD-Test* and *Legit.Health-AD-FPK-IVI* for testing. We followed a 6-fold cross-validation strategy in order to train the models. The models trained on the different folds were tested on both test sets and the results were averaged over the folds in order to reduce the variance and bias.

The second experiment was built to better understand the performance of the network on dark skin when including a tiny fraction of dark skinned patient images in the training set. In this experiment, we used *Legit.Health-AD*, *Legit.Health-AD-Test* and a subset of *Legit.Health-AD-FPK-IVI* for training and the rest for testing. The training and test subsets of *Legit.Health-AD-FPK-IVI* were obtained with a 3-fold cross-validation strategy. This means that the training set was composed of 971 light skinned patient images, *Legit.Health-AD* and *Legit.Health-AD-Test* combined, and 75 dark skinned patient images, which is a tiny fraction of the total images (8%). The dark skin test set was composed of the remaining 37 images. This split was done 3 times (3-fold) including different images in the training and test set, in order to obtain more reliable results.

In the case of visual sign severity assessment, we also ran experiments to find the optimal range, testing [0, 3], [0, 10] and [0, 100] ranges. In addition, we tested the mean and the median as the statistical measure for obtaining the ground truth of the training set. This project was entirely run on a single NVIDIA Tesla V100 (32GB) GPU.

CADx system

With the objective of making the algorithms accessible to the healthcare professional, we created a fully integrated CADx system, a web application that integrates the *Legit.Health-SCORADNet* algorithm and calculates the patient-based ASCORAD using clinical images. The CADx system includes three stages: uploading the images of the affected areas, processing the images, and reporting the Automatic Scoring of Atopic Dermatitis (ASCORAD).

In the first stage, images of affected areas are uploaded to the system using a simple user interface, depicted in Figure 3a. The user has to choose the body zone from the options defined in the original SCORAD (Stalder et al., 1993): head and neck, right upper limbs, left upper limbs, right lower limbs, left lower limbs, anterior trunk, back and genitals. In some cases, like children before the age of 2 with all bodies affected, a full body photograph can also be uploaded. In addition, the patient answers a simple questionnaire of two items, itchiness (0-10) and sleeplessness (0-10).

In the second stage, the *Legit.Health-SCORADNet* algorithm processes the images and automatically calculates the severity of atopic dermatitis by calculating the intensity of each visual sign and the surface of the lesion. Finally, the output of the algorithm is shown in an user-friendly report containing an image with the estimated lesion surface and a chart with the evolution of the ASCORAD over time. The final report of the proposed CADx system is depicted in Figure 3b.

Computing the ASCORAD requires calculating the proportion of skin covered by the lesion. We solved this by including a small piece of hardware called AI Marker, a sticker with several shapes and colors that helps translate pixels into a metric unit of measurement. The AI Marker should be kept close to the lesion and it is automatically detected. In addition, the Body Surface Area (BSA) is calculated with the patient's height and weight using the Mosteller (Lee et al., 2008)(Orimadegun and AO, 2014) formula. Once the surface of the lesion and BSA are estimated, the percentage can be calculated by dividing the surface of the lesion by the BSA (Equation 2). This allows the CADx system to calculate the final value of ASCORAD. When the AI Marker is not used, lesion surface percentage is input by the user manually, although the CADx system is still capable of calculating the visual sign intensity values automatically.

When more than one image is uploaded, the surface of the images is summed and the maximum (Dirschka et al., 2017) of each visual sign intensity is used for the ASCORAD calculation. Therefore, the final formula for N images of the whole body can be written as follows:

$$ASCORAD = \frac{1}{5} \sum_{i=1}^N \frac{S_i}{BSA} + \frac{7}{2} \sum_{i=1}^6 \max(S_i) + S_{input}$$

(2)

where S_i stands for the lesion surface in a metric unit of measurement, $S_i \in [0,3]$ stands for visual sign intensity, $S_{input} \in [0,20]$ stands for the sum of the symptoms input by the patient.

Software and statistical analysis

The models were implemented and trained using Pytorch (Paszke et al., 2019), Metrics and KFold were calculated in Python using the SciKit-Learn package (Pedregosa et al., 2012) and plotted using Matplotlib (Hunter, 2017).

RESULTS

Annotation

Evaluation of the variability of the expert dermatologist annotations in all the examined datasets was calculated. These results provided background for comparing with the results of *Legit.Health-SCORADNet*. We found out that the lesion segmentation annotation was very consistent across datasets, with an accuracy of (81.0-91.3)%, AUC of 0.91, F1 of 0.86-0.91 and RSD of (8.6-9.1)%. It can also be seen that *Legit.Health-AD-FPK-IVI* had the largest disagreement if we look at the IoU metric, with 0.80 against 0.86 and 0.91 on light skin datasets. Note that the F1 score is also the lowest for the dark skin dataset. In regards to visual sign severity assessment, *Legit.Health-AD* had more disagreement among the specialists, but the other datasets had more positive skewed distributions, meaning that the majority of the intensity values were close to 0.

Lesion surface segmentation

We compared the difference at pixel level as there was no physical reference on the images to obtain the real size of the lesions. As can be seen in Table 2, the annotations of the three datasets had a RSD close to 9%, Cohen's kappa of 0.79 and AUC around 0.90. Despite the similar results on the previously mentioned metrics, *Legit.Health-AD-FPK-IVI* seemed to have more discrepancies among the annotators, as it showed the lowest IoU and F1 values, 0.80 and 0.86, respectively.

Visual sign severity assessment

The results presented in Table 3, 4 and 5 provide background for comparing with the results of *Legit.Health-SCORADNet* in the visual sign severity assessment task. All the values are below

random RSD and above random FAR, PAR1 and PAR2 for all visual signs. Erythema is the visual sign that obtains the best Cohen's kappa value in general, and lichenification (0.06) in *Legit.Health-AD* and excoriations (0.08) and dryness (0.09) in *Legit.Health-AD-FPK-IVI* get values very close to 0. The 6 visual signs constitute a maximum of 63 points of the SCORAD, as the sum of the intensities is multiplied by $\frac{7}{5}$ (Equation 2). Giving the RSD results in terms of SCORAD points, the variability of *Legit.Health-AD* is around 11 points ($\widehat{\sigma} = 17\%$), and both *Legit.Health-AD-Test* and *Legit.Health-AD-FPK-IVI* have the same variability, on average, of 8 points ($\widehat{\sigma} = 12\%$).

Legit.Health-SCORADNet

Legit.Health-SCORADNet was validated on two experiments in which the network was trained on several data splits, as we applied a k-fold cross validation technique, 6-fold for the first experiment and 3-fold for the second experiment. All the results gathered in Table 6, 7, 8, 9 and 10 were obtained by averaging the results of the network's performance on the different data splits and are measured using the same metrics as the annotation, with the purpose of making a direct comparison of both.

Legit.Health-SCORADNet showed a good performance at visual sign severity assessment, obtaining a RMAE of 13.0% and AUC of 0.93 at surface estimation. The total execution time of *Legit.Health-SCORADNet* for a single image was 0.34s, running on an Intel Xeon Platinum 8260 CPU @ 2.40GHz.

Lesion surface segmentation

Legit.Health-SCORADNet's lesion surface segmentation results are gathered in Table 6 and 7. The AUC, IoU and F1 for light skin were 0.93, 0.64 and 0.75 respectively, whilst the results on those

metrics were 0.83, 0.32 and 0.42 for dark skin. However, when training in a small subset of dark skin images (experiment 2), the results significantly improved (0.41 on IoU and 0.33 on F1), as can be seen in Table 7. Figure 4 and Figure 5 show the ground truth and prediction for a sample case of *Legit.Health-AD-Test* and *Legit.Health-AD-FPK-IVI* respectively.

Visual sign severity assessment

As can be seen, on average, training with a ground truth obtained by applying the median and normalized in the [0,100] range obtained the best performance (Table 8). Using that configuration, we ran experiment 1 and 2 and we got a RMAE of 13.0% in *Legit.Health-AD-Test*, which had an inter-observer RMAE of 10.6%, having trained *Legit.Health-SCORADNet* on a dataset with 15.8% RMAE (Table 9). The RMAE on *Legit.Health-AD-FPK-IVI* was slightly higher, 14.3% (Table 10), when including dark skin images in the training set and 19.8%, without including dark skin images. The visual sign with the worst performance on light skin was oozing (19.4%), followed by edema (16.0%). Lichenification (19.8%) and dryness (19.3%) were the most difficult visual signs for the algorithm to correctly predict on dark skin, with edema (15.4%) also having a value above the average. Interestingly, oozing got a much lower RMAE on *Legit.Health-AD-FPK-IVI*, while both test datasets had the same oozing intensity distribution.

The distribution of predicted intensity values was plotted next to the ground truth distributions (Figure 6) to show that *Legit.Health-SCORADNet* was able to predict values in the whole range and not only the mean of the distribution.

DISCUSSION

The study shows the potential of the evaluated technology. ASCORAD shows promise as an automatic scoring system that might enable a more objective and quick evaluation. However, additional validation studies are needed in real-world settings and with diverse populations to ensure generalizability.

To put our results into clinical context, the annotated lesion area was compared with the algorithm-predicted area. As some photographs do not show the complete lesion area, live assessment cannot be directly compared with the photograph assessment methods. However, image-based area assessment by an expert and predicted area have the same basis for their analysis and are therefore directly comparable. *Legit.Health-SCORADNet* resulted in a good overall RMAE of 13.0% and an excellent AUC of 0.93 and IoU of 0.75 for lesion surface estimation on light skin.

Legit.Health-AD-Test and *Legit.Health-AD-FPK-IVI* datasets have strong positive skewed distributions for all the visual signs, which means that the most frequent intensities are 0 and 1. It seems that a vast majority of images are of mild atopic dermatitis or that the observers had a strong bias towards low intensity values. If the majority of the visual signs are close to zero intensity, it is possible that the RSD reflected lower disagreement (9% versus 17% in *Legit.Health-AD*). In fact, Oranje et al. (Oranje et al., 2007) (RSD 20%) found an RSD of 20%, very close to the inter-observer variability found in *Legit.Health-AD*.

Looking at Cohen's Kappa values it seems that some of the visual signs like lichenification in *Legit.Health-AD* and excoriations and dryness in *Legit.Health-AD-FPK-IVI* have a null inter-observer agreement. However, Cohen's Kappa is a statistical measure for nominal classification problems and metrics like RSD, MAE, FAR, PAR 1 and PAR 2 show that the annotation of the

specialists is far from random. For example, the visual sign excoriations in *Legit.Health-AD-FPK-IVI* obtains a Cohen's Kappa value of 0.09 and PAR 2 of 95.5%, far from the random value (62%).

In short, we have proved that a CNN trained with the observer's average results can achieve a similar RMAE to the one of the experts. Furthermore, our automatic method outputs a value in the range [0,100] for each visual sign instead of [0,3] as the usual SCORAD, broadening the spectrum of possible outputs and turning the discrete problem into more continuous.

CONCLUSION

A deep learning algorithm could simplify the assessment of atopic dermatitis, a very common skin disease that affects 15-20% of children (Asher et al., 2006) and 1-3% of adults worldwide. Scoring systems like SCORAD and EASI have an inter-observer variability and are time consuming. An AI automated approach like ours may help to reduce such bias and therefore be a more precise and objective criterion for evaluation in pharmaceutical studies. Our results show that deep learning may be noticed as a fast and objective alternative method for the automatic assessment of atopic dermatitis with great potential, already achieving results comparable to human expert assessment, whilst missing inter-observer variability and being more time efficient. ASCORAD could also be used in situations where face-to-face consultations are not possible, providing an automatic assessment of clinical signs and lesion surface. It could also be a potential tool to reduce the time and effort of training clinical assessors for clinical trials and in clinical practice.

Further work needs to be done in order to prove the validity, responsiveness and reliability of the system in real-world clinical practise. Despite the dataset used in this study captures the variability of a wide range of parameters, the algorithm should be tested on other datasets to prove its robustness and generalizability, in particular to dark skin tones.

In the future, we intend to test ASCORAD in validation studies in which the objective part of the SCORAD will be assessed in person by the dermatologist. Comparing the result of the algorithm to face-to-face assessment is crucial because some visual signs like oedema, dryness or oozing might present more difficulties in estimating the severity via image than in person. Furthermore, the AI Marker will be used in this study, helping the CADx system correctly estimate the surface by converting lesion pixels into a metric unit of measurement.

The process of taking the images does not necessarily have to be done by the physician themselves, we believe that our algorithm has the potential to reduce costs in dermatology by saving time, whilst improving documentation of the evolution of the disease. This could also be interesting for the application in pharmaceutical clinical trials.

LIMITATIONS

Legit.Health datasets used in this study have a low number of images to be considered robust and, therefore, a larger number of images will be needed in future studies in order to obtain more statistically significant results. As most of the images used in this study are of light skin (Fitzpatrick's I, II and III), evidence of the algorithm's performance on dark skin (Fitzpatrick's VI, V and VI) is limited. The results of the algorithms trained on light skin and tested on dark skin were significantly worse, however, when training with a small proportion (8%) of dark skin images, the results improved a lot (41% on IoU and 33% on F1). These results indicate that it might be possible to create a single algorithm for all the skin types, so further work will be focused on developing a system that can also assist doctors in the assessment of atopic dermatitis on darker skin types, reducing healthcare disparities in skin of color (Adamson and Smith, 2018). In addition, *Legit.Health-AD-Test* and *Legit.Health-AD-FPK-IVI* lack a detailed demographic

characterization. Gathering a complete and detailed demographic data will be essential in future studies in order to identify potential sources of bias in the datasets (Daneshjou et al., 2021).

Another limitation that needs to be discussed is that to accurately estimate the surface of the lesion, the AI Marker must be used as a reference. This small piece of hardware is a sticker that contains a wide range of colors and shapes, which allows translating the lesion surface from pixels to a metric unit of measurement. If the AI Marker does not appear next to the lesion, or does not appear in the photo, the algorithm cannot calculate the final ASCORAD.

An additional limitation to take into account is the total time of ASCORAD, from taking the picture or pictures to getting the final results. This was not measured in the study and depends on many factors such as the number and location of the lesions. It is also important to note that the time is reduced to zero if the photos are taken at home by the patients in a fully remote follow-up consultation. However, this might create some additional problems like getting poor quality images or difficulties for the patients to access some body zones by themselves. We will address these topics in future studies.

DATA AVAILABILITY STATEMENT

The images of *Legit.Health-AD*, *Legit.Health-AD-Test* and *Legit.Health-AD-FPK-IVI* datasets related to this article can be found at <http://www.atlasdermatologico.com.br/>, hosted at Dermatology Atlas, <http://www.danderm-pdv.is.kkh.dk/>, hosted at Danderm, <https://www.dermatlas.net/>, hosted at Interactive Dermatology Atlas, <https://www.dermis.net/dermisroot/en/home/index.htm>, hosted at DermIS (Diepgen and Eysenbach, 1998), <https://dermnetnz.org/>, hosted at DermNet NZ and <http://www.hellenicdermatlas.com/en/>, hosted at Hellenic Dermatological Atlas.

ORCIDs

Alfonso Medela: <https://orcid.org/0000-0001-5859-5439>

Taig Mac Carthy: <https://orcid.org/0000-0001-5583-5273>

S. Andy Aguilar Robles: <https://orcid.org/0000-0003-0618-6179>

Carlos M. Chiesa-Estomba: <https://orcid.org/0000-0001-9454-9464>

Ramon Grimalt: <https://orcid.org/0000-0001-7204-8626>

CONFLICT OF INTEREST

The authors state no conflict of interest.

ACKNOWLEDGMENTS

The authors thank Dr. Fernando Alfageme Roldán for technical advice and IBM for providing the computing infrastructure for the deep learning experiments.

AUTHOR CONTRIBUTIONS STATEMENT

Conceptualization: AM, RG; Data Curation: AM; Formal Analysis: AM, CMCE; Investigation: AM, TMC, AA, CMCE, RG; Methodology: AM, TMC, RG; Project Administration: AA; Visualization: TMC; Writing - Original Draft Preparation: AM; Writing - Review and Editing: AM, TMC, AA, CMCE, RG

REFERENCES

- Adamson, A. S. & Smith, A. (2018). Machine Learning and Health Care Disparities in Dermatology. *JAMA dermatology*, 154(11), 1247–1248.
<https://doi.org/10.1001/jamadermatol.2018.2348>
- Asher, M. I., Montefort, S., Björkstén, B., Lai, C. K., Strachan, D. P., Weiland, S. K. et al. (2006). Worldwide time trends in the prevalence of symptoms of asthma, allergic rhinoconjunctivitis and eczema in childhood: Isaac phases one and three repeat multicountry cross-sectional survey. *Lancet* 368, 733–43, DOI: 10.1016/S0140-6736(06)69283-0
- Berke, R., Singh, A. & Guralnick, M. (2012). Atopic dermatitis: An overview. *Am. family physician* 86, 35–42
- Bieber, T. (2008). Atopic dermatitis. *The New Engl. journal medicine* 358, 1483—1494, DOI: 10.1056/nejmra074081
- Butler, É., Lundqvist, C. & Axelsson, J. (2020). *Lactobacillus reuteri* dsm 17938 as a novel topical cosmetic ingredient: A proof of concept clinical study in adults with atopic dermatitis. *Microorganisms* 8
- Chopra, R., Vakharia, P. P., Sacotte, R., Patel, N., Immaneni, S., White, T. et al (2017). Relationship between easi and scorad severity assessments for atopic dermatitis. *J. Allergy Clin. Immunol.* 140, DOI: 10.1016/j.jaci.2017.04.052
- Daneshjou, R., Smith, M. P., Sun, M. D., Rotemberg, V., & Zou, J. (2021). Lack of Transparency and Potential Bias in Artificial Intelligence Data Sets and Algorithms: A Scoping

Review. *JAMA dermatology*, 157(11), 1362–1369.

<https://doi.org/10.1001/jamadermatol.2021.3129>

Dash, M., Londhe, N., Ghosh, S., Raj, R. & Sonawane, R. (2020). A cascaded deep convolution neural network based cadx system for psoriasis lesion segmentation and severity assessment.

Appl. Soft Comput. 91, 106240

Dash, M., Londhe, N., Ghosh, S., Semwal, A. & Sonawane, R. (2019). Pslsnet: Automated psoriasis skin lesion segmentation using modified u-net-based fully convolutional network.

Biomed. Signal Process. Control. 52, 226–237

Diepgen, T. L., & Eysenbach, G. (1998). Digital images in dermatology and the Dermatology Online Atlas on the World Wide Web. *The Journal of dermatology*, 25(12), 782–787.

<https://doi.org/10.1111/j.1346-8138.1998.tb02505.x>

Dirschka, T., Pellacani, G., Micali, G., Malvehy, J., Stratigos, A. J., Casari, A. et al. (2017). A proposed scoring system for assessing the severity of actinic keratosis on the head: actinic keratosis area and severity index. *Journal of the European Academy of Dermatology and*

Venereology : JEADV, 31(8), 1295–1302. <https://doi.org/10.1111/jdv.14267>

Drucker, A., Wang, A, Li, W., Severson, E., Block, J. & Qureshi, A. (2017). The Burden of Atopic Dermatitis: Summary of a Report for the National Eczema Association. *Journal of Investigative Dermatology*. 137.10.1016/j.jid.2016.07.012

Eichenfield, L., Tom, W., Chamlin, S., Feldman, S., Hanifin, J., Simpson, E. et al. (2014).

Guidelines of care for the management of atopic dermatitis: Section 1. Diagnosis and assessment

of atopic dermatitis Work Group. *Journal of the American Academy of Dermatology*. 70.

10.1016/j.jaad.2013.10.010

Guimarães, P., Batista, A., Zieger, M., Kaatz, M. & Koenig, K. (2020). Artificial intelligence in multiphoton tomography: Atopic dermatitis diagnosis. *Sci. Reports* 10, DOI: 10.1038/s41598-

020-64937-x

Gustafson, E., Pacheco, J., Wehbe, F., Silverberg, J. & Thompson, W. (2017). A machine learning algorithm for identifying atopic dermatitis in adults from electronic health records. vol.

2017, 83–90, DOI: 10.1109/ICHI.2017.31

Hanifin, J. M., Thurston, M., Omoto, M., Cherill, R., Tofte, S. J., & Graeber, M. (2001). The eczema area and severity index (easi): assessment of reliability in atopic dermatitis. *Exp.*

Dermatol. 10

He, K., Zhang, X., Ren, S. & Sun, J. (2015). Deep residual learning for image recognition.

1512.03385

Hunter, J. D. (2007). Matplotlib: A 2d graphics environment. *Comput. Sci. Eng.* 9, 90–95

Lakshminarayanan, B. & Teh, Y. W (2013). Inferring ground truth from multi-annotator ordinal data: a probabilistic approach. 1305.0015

Lee, J.Y., Choi, J.W. & Kim, H (2008). Determination of body surface area and formulas to estimate body surface area using the alginate method. *J Physiol Anthropol.* DOI:

10.2114/jpa2.27.71. PMID: 18379164

Nahm, D. H., Ye, Y. M., Shin, Y. S., Park, H. S., Kim, M. E., Kwon, B. et al. (2020). Efficacy, Safety, and Immunomodulatory Effect of the Intramuscular Administration of Autologous Total

Immunoglobulin G for Atopic Dermatitis: A Randomized Clinical Trial. *Allergy, asthma & immunology research*, 12(6), 949–963. <https://doi.org/10.4168/aair.2020.12.6.949>

Nutten, S. (2015). Atopic dermatitis: global epidemiology and risk factors. *Annals nutrition amp; metabolism* 66 Suppl 1, 8—16, DOI: 10.1159/000370220

Oranje, A., Glazenburg, E., Wolkerstorfer, A. & De Waard-van der Spek, F. (2007). Practical issues on interpretation of scoring atopic dermatitis: The scorad index, objective scorad and the three-item severity score. *The Br. journal dermatology* 157, 645–8, DOI: 10.1111/j.1365-2133.2007.08112.x

Orimadegun, A. & AO, O. (2014). Evaluation of five formulae for estimating body surface area of nigerian children. *Ann Med*

Heal. Sci Res 4, 889–898, DOI: 10.4103/2141-9248.144907.

Pal, A., Chaturvedi, A., Garain, U., Chandra, A. & Chatterjee, R. (2016). Severity grading of psoriatic plaques using deep cnn based multi-task learning. 2016 23rd Int. Conf. on Pattern Recognit. (ICPR) 1478–1483

Pan, S. J. & Yang, Q. (2010). A survey on transfer learning. *IEEE Transactions on Knowl. Data Eng.* 22, 1345–1359

Panahi Y., Davoudi S.M., Sahebkar A., Beiraghdar F., Dadjo Y., Feizi, I. et al. (2011). Efficacy of aloe vera/olive oil cream versus betamethasone cream for chronic skin lesions following sulfur mustard exposure: A randomized double-blind clinical trial. *Cutan. ocular toxicology* 31, 95–103, DOI: 10.3109/ 15569527.2011.614669

- Paszke, A., Gross, S., Massa, F., Lerer, A., Bradbury, J., Chanan, G. et al. (2019). Pytorch: An imperative style, high-performance deep learning library. In *Advances in neural information processing systems*, 8026–8037
- Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O. et al. (2012). Scikit-learn: Machine learning in python. *CoRR abs/1201.0490*. 1201.0490
- Ronneberger, O., Fischer, P. & Brox, T. (2015). U-net: Convolutional networks for biomedical image segmentation. *CoRR abs/1505.04597*. 1505.04597
- Rothe, R., Timofte, R. & Gool, L. (2015). Dex: Deep expectation of apparent age from a single image. *2015 IEEE Int. Conf. on Comput. Vis. Work. (ICCVW)* 252–257
- Russakovsky, O., Deng, J., Su, H., Krause, J., Satheesh, S., Ma, S. et al. (2014). Imagenet large scale visual recognition challenge. *CoRR abs/1409.0575*. 1409.0575
- Schmitt J., Langan, S., Deckert, S., Svensson, A., von Kobyletzki, L., Thomas, K. & Spuls, P. (2013). Assessment of clinical signs of atopic dermatitis: A systematic review and recommendation. *J. Allergy Clin. Immunol.* 132, 1337–1347, DOI: 10.1016/j.jaci.2013.07.008)
- Silverberg J. I., Simpson E. L., Thyssen J. P., Gooderham M., Chan G., Feeney C. et al. (2020). Efficacy and safety of abrocitinib in patients with moderate-to-severe atopic dermatitis: A randomized clinical trial. *JAMA Dermatol.* 156, DOI: 10.1001/jamadermatol.2020.1406
- Stalder, J. F., Barbarot, S., Wollenberg, A., Holm, E. A., De Raeve, L., Seidenari, S. et al (2011). Patient oriented scorad (po-scorad): a new self assessment scale in atopic dermatitis, validated in europe. *Allergy* 66, 1114–21, DOI: 10.1111/j.1398-9995.2011.02577.x

Stalder, J. F., Taïeb, A., Atherton, D. J., P. Bieber, P., Bonifazi, E., Broberg, A. et al. (1993). Severity scoring of atopic dermatitis: The scorad index: Consensus report of the european task force on atopic dermatitis. *Dermatology* 186, 23–31, DOI: 10.1159/000247298

Tan, M. & Le, Q. V. (2019). Efficientnet: Rethinking model scaling for convolutional neural networks. *CoRR abs/1905.11946*. 1905.11946

Wu, H., Yin, H., Chen, H., Sun, M., Liu, X., Yu, Y. et al. (2020). A deep learning, image based approach for automated diagnosis for inflammatory skin diseases. *Annals of translational medicine*, 8(9), 581. <https://doi.org/10.21037/atm.2020.04.39>

D. Xu, Y. Shi, I. W. Tsang, Y. -S. Ong, C. Gong and X. Shen (2020). "Survey on Multi-Output Learning," in *IEEE Transactions on Neural Networks and Learning Systems*, vol. 31, no. 7, pp. 2409-2429, July 2020, doi: 10.1109/TNNLS.2019.2945133

Yoo J., Choi J. Y., Lee B. Y., Shin C. H., Shin J. W., Huh C. H., Na J. I. et al (2020). Therapeutic effects of saline groundwater solution baths on atopic dermatitis: A pilot study. *Evidence-Based Complementary Altern. Medicine* 2020, 1–5, DOI: 10.1155/2020/8303716

TABLES

Table 1. Demographic characteristics.

Dataset	Age groups (%)						Sex (%)		Skin type (%)	
	<18	18-29	30-39	40-49	50-64	>65	Male	Female	Light	Dark
<i>Legit.Health-AD</i>	31	23	26	14	4	2	39	61	100	0
<i>Legit.Health-AD-Test</i>	-	-	-	-	-	-	-	-	100	0
<i>Legit.Health-AD-FPK-IVI</i>	-	-	-	-	-	-	-	-	0	100

Table 2. Annotator's performance in lesion surface segmentation. These results provide background for comparing with the results of *Legit.Health-SCORADNet*.

Dataset	ACC	AUC	IoU	F1	RSD	Cohen's kappa
<i>Legit.Health-AD</i>	86.9	0.91	0.91	0.88	8.6	0.78
<i>Legit.Health-AD-Test</i>	81.0	0.91	0.86	0.91	9.1	0.79
<i>Legit.Health-AD-FPK-IVI</i>	91.3	0.91	0.80	0.86	9.0	0.80

Abbreviations: ACC — Accuracy, AUC — Area Under the Curve, IoU — Intersection over Union, F1 — F1 Score, RSD — Relative Standard Deviation.

Table 3. Annotator's performance in *Legit.Health-AD* visual sign severity assessment. These results provide background for comparing with the results of *Legit.Health-SCORADNet*.

Visual sign	RSD	RMAE (mean)	RMAE (median)	FAR	PAR 1	PAR 2	Cohen's kappa
Erythema	11.5	10.7	8.3	33.1	92.0	94.5	0.34

Edema	16.2	14.7	11.9	21.3	74.1	84.9	0.15
Oozing	20.0	18.2	14.8	18.0	59.6	79.3	0.19
Excoriations	17.4	15.9	12.9	22.6	66.5	81.2	0.17
Lichenification	20.3	18.3	15.1	10.7	59.1	74.6	0.06
Dryness	18.7	16.9	12.8	20.0	69.3	82.3	0.14
Average	17.4	15.8	13.8	14.4	64.7	79.3	0.17

Abbreviations: RSD — Relative Standard Deviation, RMAE— Relative Mean Absolute Error, FAR — Full Agreement Rate, PAR — Partial Agreement Rate.

Table 4. Annotator’s performance in *Legit.Health-AD-Test* visual sign severity assessment.

These results provide background for comparing with the results of *Legit.Health-SCORADNet*.

Visual sign	RSD	RMAE (mean)	RMAE (median)	FAR	PAR 1	PAR 2	Cohen's kappa
Erythema	12.1	11.2	8.8	34.0	88.0	91.5	0.35
Edema	7.9	7.3	5.6	55.8	93.1	96.7	0.22
Oozing	10.3	9.5	7.5	44.4	89.9	93.1	0.39
Excoriations	12.7	11.6	9.4	39.7	79.0	87.1	0.20
Lichenification	10.1	9.3	7.4	46.8	88.0	92.9	0.21
Dryness	16.5	14.9	12.2	20.4	72.4	80.3	0.19
Average	11.6	10.6	8.5	40.2	85.0	90.3	0.26

Abbreviations: RSD — Relative Standard Deviation, RMAE — Relative Mean Absolute Error, FAR — Full Agreement Rate, PAR — Partial Agreement Rate.

Table 5. Annotator’s performance in *Legit.Health-AD-FPK-IVI* visual sign severity assessment.

These results provide background for comparing with the results of *Legit.Health-SCORADNet*.

Visual sign	RSD	RMAE (mean)	RMAE (median)	FAR	PAR 1	PAR 2	Cohen's kappa
Erythema	11.9	10.8	8.8	42.3	80.1	88.2	0.23
Edema	8.6	8.0	6.3	54.0	90.9	94.5	0.13
Oozing	12.7	11.6	9.4	35.1	81.9	87.3	0.27
Excoriations	9.7	9.0	7.0	45.0	92.7	95.5	0.08
Lichenification	13.3	12.2	9.7	27.9	85.5	90.9	0.27
Dryness	18.2	16.4	13.4	10.8	70.2	81.0	0.09
Average	12.4	11.3	9.1	35.9	86.6	89.6	0.18

Abbreviations: RSD — Relative Standard Deviation, RMAE — Relative Mean Absolute Error, FAR — Full Agreement Rate, PAR — Partial Agreement Rate.

Table 6. *Legit.Health-SCORADNet*'s results in light skin lesion surface segmentation.

	ACC (%) - 95% CI	AUC - 95% CI	IoU - 95% CI	F1 - 95% CI
Lesion surface	84.6 (80.9-88.3)	0.93 (0.90-0.96)	0.64 (0.59-0.69)	0.75 (0.71-0.79)

Abbreviations: ACC — Accuracy, AUC — Area Under the Curve, IoU — Intersection over Union, F1 — F1 Score, CI — Confidence Interval.

Table 7. *Legit.Health-SCORADNet*'s results in dark skin lesion surface segmentation.

Results are divided by experiment. Experiment 1 algorithm was trained on purely light skinned patient images and experiment 2 algorithm was trained on mixed data containing an 8% of dark skinned patient images.

	Experiment 1				Experiment 2			
	ACC (%) -	AUC -	IoU - 95%	F1 - 95%	ACC (%) -	AUC -	IoU - 95%	F1 - 95%
	95% CI	95% CI	CI	CI	95% CI	95% CI	CI	CI
Lesion surface	74.0 (65.9-82.1)	0.83 (0.76-0.90)	0.32 (0.23-0.41)	0.42 (0.33-0.51)	79.2 (66.3-92.1)	0.87 (0.76-0.98)	0.45 (0.29-0.61)	0.55 (0.39-0.71)

Abbreviations: ACC — Accuracy, AUC — Area Under the Curve, IoU — Intersection over Union, F1 — F1 Score, CI — Confidence Interval.

Table 8. *Legit.Health-SCORADNet*'s results in visual sign severity assessment. The models were trained on *Legit.Health-AD* using a different range and ground truth method and tested on *Legit.Health-AD-Test* and *Legit.Health-AD-FPK-IVI*.

		<i>Legit.Health-AD-Test</i>		<i>Legit.Health-AD-FPK-IVI</i>	
Range	Training GT	RMAE \square^1 - 95%	RMAE \square^2 - 95%	RMAE 1 - 95%	RMAE 2 - 95%
		CI	CI	CI	CI
[0, 3]	Median	13.6 (9.7-17.5)	14.3 (10.4-18.2)	21.2 (17.3-25.0)	20.8 (16.9-24.7)
[0, 10]	Median	14.3 (10.4-18.2)	13.2 (9.3-17.0)	22.8 (18.9-26.7)	20.0 (16.0-23.9)
[0, 100]	Median	14.4 (10.5-18.3)	13.0 (9.1-16.9)	22.6 (18.7-26.5)	19.8 (15.9-23.7)

[0, 100]	Mean	13.5 (9.6-17.4)	13.4 (9.5-17.3)	21.1 (17.2-25.0)	19.9 (16.0-23.8)
----------	------	-----------------	-----------------	------------------	------------------

Abbreviations: RMAE — Relative Mean Absolute Error, CI — Confidence Interval.

¹RMAE 1 is obtained by applying the argmax function to the prediction.

²RMAE 2 is obtained by applying the DEX method to the prediction.

Table 9. *Legit.Health-SCORADNet's* results in light skin visual sign severity assessment.

Visual sign	RMAE \square^1 - 95% CI	RMAE \square^2 - 95% CI
Erythema	14.1 (10.2-18.0)	13.3 (9.4-17.2)
Edema	16.1 (12.2-20.0)	16.0 (12.1-19.9)
Oozing	22.3 (18.4-26.2)	19.4 (15.5-23.3)
Excoriations	11.5 (7.6-15.4)	9.6 (5.7-15.4)
Lichenification	10.3 (6.4-14.2)	8.7 (4.8-12.6)
Dryness	12.4 (8.5-16.3)	11.3 (7.4-15.2)
Average	14.4 (10.5-18.3)	13.0 (9.1-16.9)

Abbreviations: RMAE — Relative Mean Absolute Error, CI — Confidence Interval.

¹RMAE 1 is obtained by applying the argmax function to the prediction.

²RMAE 2 is obtained by applying the DEX method to the prediction.

Table 10. *Legit.Health-SCORADNet's* results in dark skin visual sign severity assessment.

Results are divided by experiment. Experiment 1 algorithm was trained on purely light skinned patient images and experiment 2 algorithm was trained on mixed data containing an 8% of dark skinned patient images.

Visual sign	Experiment 1		Experiment 2	
	RMAE \square^1 - 95% CI	RMAE \square^2 - 95% CI	RMAE 1 - 95% CI	RMAE 2 - 95% CI
Erythema	17.8 (13.9-21.7)	15.7 (11.8-19.6)	16.2 (12.2-20.2)	14.3 (10.3-18.3)
Edema	16.8 (12.9-20.7)	18.6 (14.7-22.5)	18.1 (14.1-22.0)	15.4 (11.4-19.4)
Oozing	24.9 (21.0-28.8)	22.7 (18.8-26.6)	9.3 (5.3-13.3)	9.0 (5.0-13.0)
Excoriations	10.1 (6.2-14.0)	9.6 (5.7-13.5)	10.2 (6.2-14.2)	8.0 (4.0-12.0)
Lichenification	25.9 (22.0-29.8)	20.6 (16.7-24.5)	24.0 (20.0-28.0)	19.8 (15.8-23.8)
Dryness	39.9 (36.0-43.8)	31.7 (27.8-35.6)	26.0 (22.0-30.0)	19.3 (15.3-23.3)
Average	22.6 (18.7-26.5)	19.8 (15.9-23.7)	17.3 (13.3-21.3)	14.3 (10.3-18.3)

Abbreviations: RMAE — Relative Mean Absolute Error, CI — Confidence Interval.

¹RMAE 1 is obtained by applying the argmax function to the prediction.

²RMAE 2 is obtained by applying the DEX method to the prediction.

FIGURE LEGENDS

Figure 1. Comparison of the intensity level distribution by visual sign of the datasets used in the study.

Figure 2. The visual signs that compose the SCORAD. Each visual sign can be classified into four intensity levels: none (0), mild (1), moderate (2) and severe (3). The multi-output EfficientNet-B0 network trained for visual sign intensity estimation has one head for each visual sign.

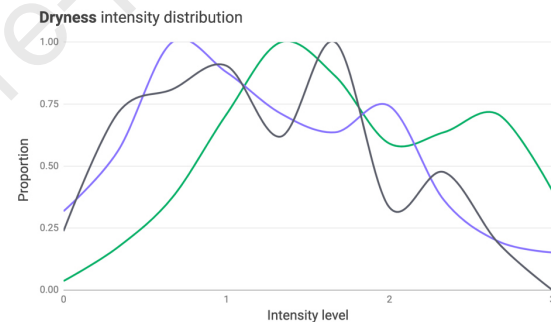
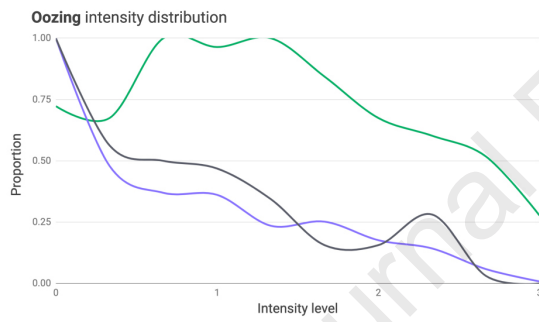
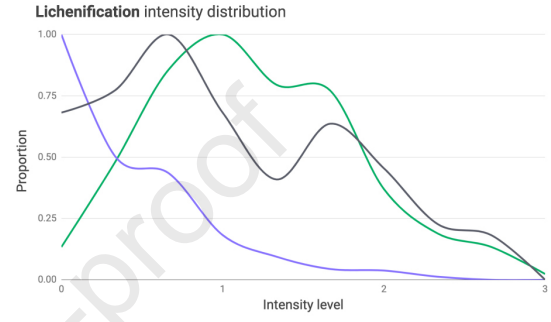
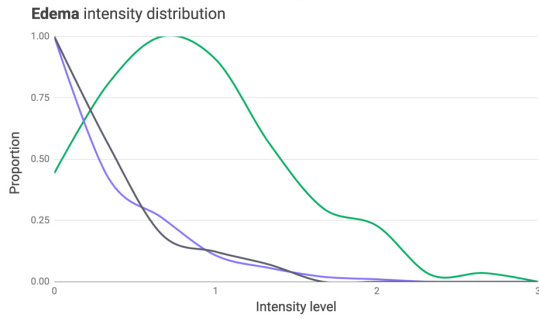
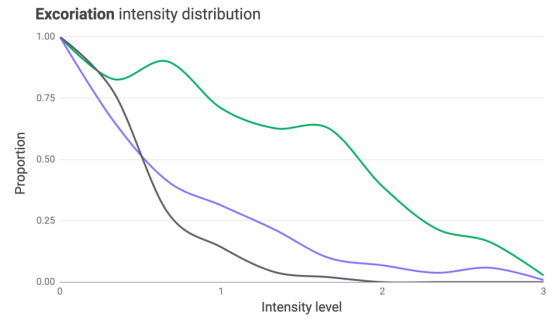
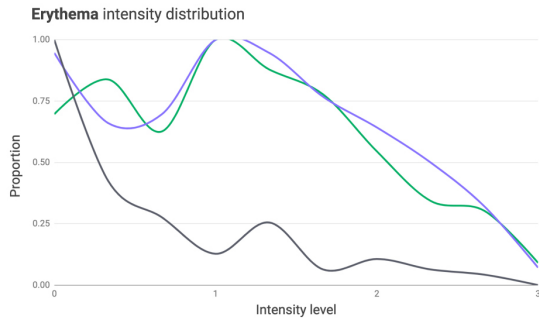
Figure 3. CADx system. a) Illustration of the questionnaire. **b)** Illustration of the report generated by the CADx system. The report contains the evolution across time of the ASCORAD, the last reported ASCORAD item by item, a picture of the lesion surface predicted by the algorithm, the final ASCORAD score with its translation to a category and some additional information like image quality.

Figure 4. Lesion surface segmentation masks. a) Original image. **b)** *Legit.Health-SCORADNet*'s prediction. **c)** Ground truth. **d)** Mask drawn by the first specialist. **e)** Mask drawn by the second specialist. **f)** Mask drawn by the third specialist. *Legit.Health-AD-Test* sample image gathered from Danderm dermatology atlas with the author's consent.

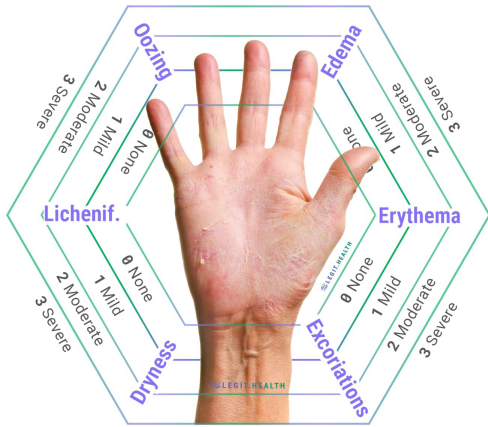
Figure 5. Results of experiment 1 and 2 models on a dark skin image. a) The predicted surface mask of the model trained on light skin. **b)** The predicted surface mask of the model trained on both light and dark skin. **c)** The ground truth mask. *Legit.Health-AD-FPK-IVI* sample image gathered from Danderm dermatology atlas with the author's consent.







Figure 6. *Legit.Health-AD-Test* visual sign intensity distribution of ground truth labels and predictions. The horizontal axis is in the range $[0, 100]$, as the results are given using the best performing model, which was trained with ground truth labels in that range.

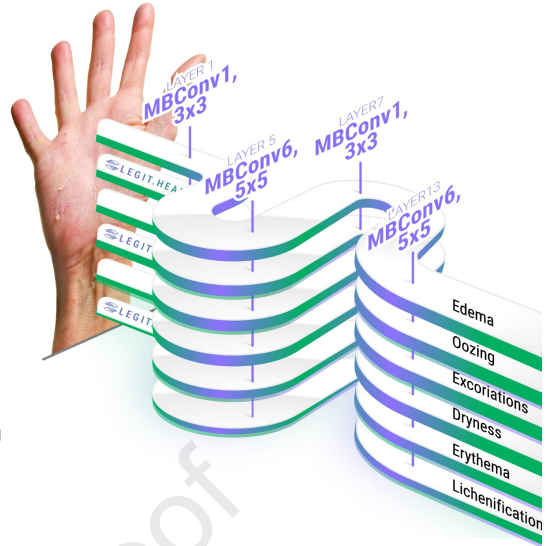
Journal Pre-proof



■ LegitHealth-AD ■ LegitHealth-AD-Test ■ LegitHealth-AD-FPK-IVI



-  Edema
-  Oozing
-  Excoriations
-  Dryness
-  Erythema
-  Lichenification



Journal Pre-proof

a

app.legit.health/

Upload photo for a new case

Select location Instructions Upload picture Questionnaire

ASCORAD Automatic Scoring of Atopic Dermatitis

Itchiness
1

Sleeplessness
3

DLQI Dermatology Life Quality Index

Back Finish

b

app.legit.health/

LEGIT.HEALTH

Dr. Ramón Grimalt

Albert Castells
Image sent by patient

Send Print

Pathology
Atopic dermatitis

Timestamp
3 Jul, 2021, 13:34 CET

Image quality (DIQA)
97%

See patient's history

Evolution across time

8 May, 2021
Score: 45

3 Jul, 2021
Score: 28

Severity
28
Moderate
Re-adjust manually

ASCORAD
Automatic SCOring Of At...

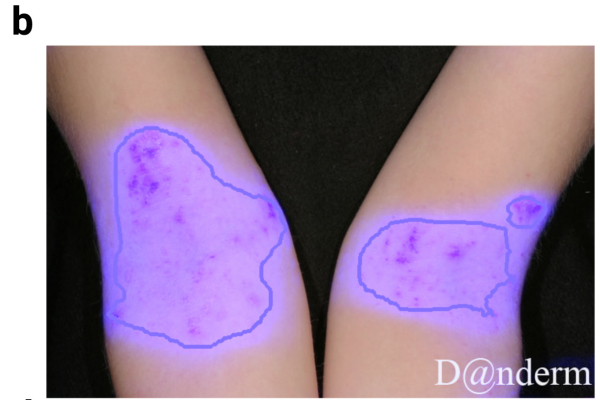
Score: 28

Erythema Mild (1)

Edema Mild (1)

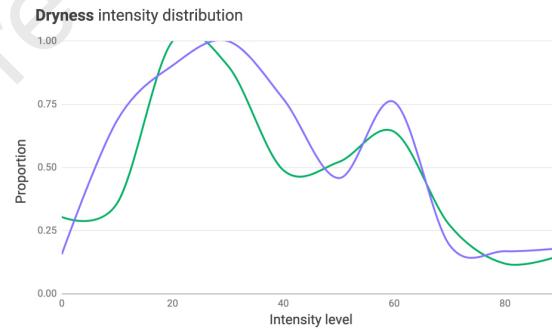
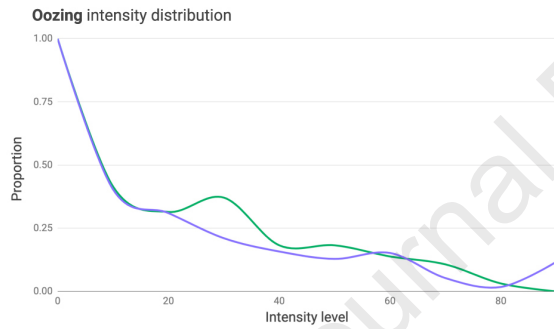
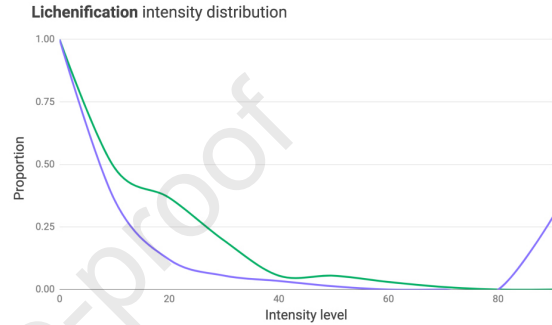
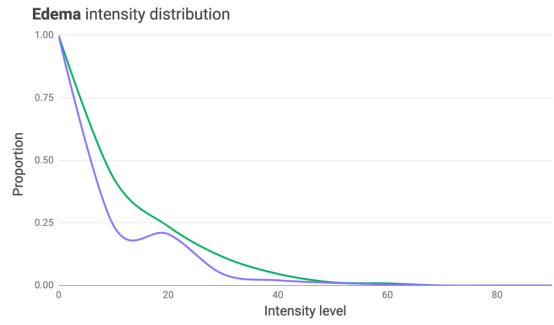
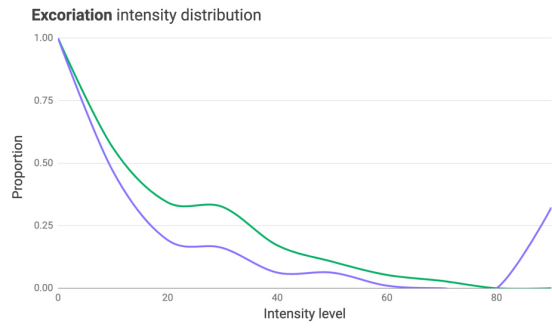
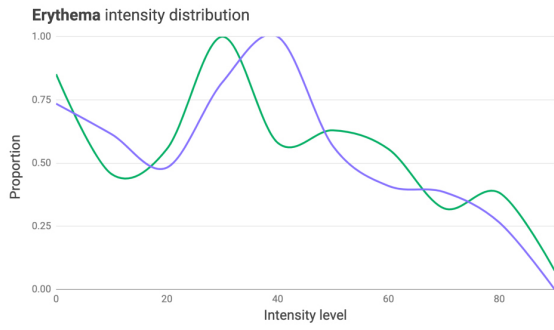
Oozing None (0)

Excoriation





Journal Pre-proof



■ Ground truth ■ Prediction