EczemaNet: Automating Detection and Severity Assessment of Atopic Dermatitis

Kevin Pan[1], Guillem Hurault[1], Kai Arulkumaran[1], Hywel C. Williams[2], and Reiko J. Tanaka[1]

[1] Imperial College London, London SW7 2AZ, UK
[2] University of Nottingham, Nottingham NG7 2UH, UK
r.tanaka@imperial.ac.uk

Abstract. Atopic dermatitis (AD), also known as eczema, is one of the most common chronic skin diseases. AD severity is primarily evaluated based on visual inspections by clinicians, but is subjective and has large inter- and intra-observer variability in many clinical study settings. To aid the standardisation and automating the evaluation of AD severity, this paper introduces a CNN computer vision pipeline, EczemaNet, that first detects areas of AD from photographs and then makes probabilistic predictions on the severity of the disease. EczemaNet combines transfer and multitask learning, ordinal classification, and ensembling over crops to make its final predictions. We test EczemaNet using a set of images acquired in a published clinical trial, and demonstrate low RMSE with well-calibrated prediction intervals. We show the effectiveness of using CNNs for non-neoplastic dermatological diseases with a medium-size dataset, and their potential for more efficiently and objectively evaluating AD severity, which has greater clinical relevance than mere classification.

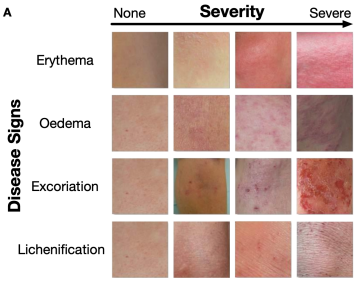
Keywords: Eczema · Multitask · Uncertainty

1 Introduction

Atopic dermatitis (or ezcema; AD) is a chronic skin disease affecting 15-30% of children and 2-10% of adults worldwide [27]. It is characterised by recurrent skin inflammation that can severely impact patients' lifestyles, with detrimental effects on social, academic, and occupational aspects of their lives. While current treatments aim to manage dynamic and unpredictable fluctuations of AD symptoms, only 24% of patients and caregivers feel confident that they can manage AD symptoms adequately [31]. Automating the evaluation of AD severity would allow us to assist research into the disease and enable patients to become more involved in the management of their condition. Remote assessment of AD symptoms by automated evaluation would enhance data-enabled efficient clinical trials by reducing the burden of parties involved and minimise detection bias in clinical trials that test interventions.

Several clinical scores are commonly used to grade the severity of AD, including the Six Area, Six Sign Atopic Dermatitis (SASSAD) score [4], the Three

Item Severity Score (TISS) [28], and the Eczema Area and Severity Index (EASI) [11], the latter of which is recommended by the Harmonising Outcome Measure for Eczema organisation [20]. Each of these are defined according to a combination of the severities of 7 disease signs[3] (Fig. 1): cracking (Cra.), dryness (Dry.), erythema (Ery.), excoriation (Exc.), exudation (Exu.), lichenification (Lic.) and oedema (Oed.). However, due to the lack of sufficient clinical training materials, and the non-intuitive nature of some disease signs (e.g., "dryness" versus "cracking"), inter- and intra-rater reliability is poor [19]. Our goal is to improve the reliability of these scoring systems through computer-aided evaluation of the different disease signs.



B

Disease Signs	Severity Score		
	SASSAD	TISS	EASI
Cracking (Cra.)	x		
Dryness (Dry.)	x		
Erythema (Ery.)	x	x	x
Excoriation (Exc.)	x	x	x
Exudation (Exu.)	x		
Lichenification (Lic.)	x		x
Oedema (Oed.)		x	x

Fig. 1: Disease signs and their relationship to severity scores. A) Examples of the 4 disease signs associated with EASI. Reproduced from [11]. B) A list of disease signs used for calculating SASSAD, TISS and EASI.

In recent years, machine-learning-based methods using convolutional neural networks (CNNs) have reached dermatologist-level performance on classifying skin cancers [7, 5]. However, due to the lack of standardised clinical datasets beyond skin cancer, applications of CNNs for non-cancerous diseases have mostly been limited to automatic disease diagnosis of skin lesions [10, 16, 29]. Whether a lesion can be attributed to AD is of limited value to already diagnosed patients, and does not address the important challenge of assessing the overall severity of the disease, whose lesions are spatially distributed over the entire body and can exhibit multiple symptoms of varying intensities.

In this paper, we introduce a novel computer vision pipeline, EczemaNet, that is capable of detecting and evaluating the severity of AD from camera images. In comparison to prior work [2], we use deep learning to learn relevant features from the data (as opposed to hand-engineered features), produce probabilistic predictions, and evaluate our method on a far larger dataset. Our pipeline uses CNNs to first detect regions-of-interest (RoI) from an image to make image crops, and then evaluate the severity of the 7 disease signs in each crop. Our input images often include background, clothes, etc. while most pipelines expect closely cropped images [26]. Similarly to recent work on psoriatic plaque severity assessment [17], we use ordinal classification to predict the severity of multiple disease signs simultaneously. However, we also propagate the uncertainties over these

[3] As well as the area of the affected region in the case of EASI.

predictions to produce a final set of severity scores (SASSAD, TISS and EASI) simultaneously, and show that using multiple crops and probabilistic predictions allows us to make well-calibrated predictions with low root mean squared error (RMSE). These properties make EzcemaNet a promising proof-of-concept for the use of CNNs in clinical trials, with downstream applications in personalised therapies for AD.

2 Data

Our data originates from the Softened Water Eczema Trial (SWET), which is a randomised controlled trial of 12 weeks duration followed by a 4-week crossover period, for 310 AD children aged from 6 months to 16 years [25]. The original data contains 1393 photos of representative AD regions taken during their clinic visits, along with the corresponding severity of each disease sign. During each visit, a disease assessment was made for SASSAD and TISS, using the 7 disease signs labelled for each image. The severity of each sign was determined on an ordinal scale: *none* (0), *mild* (1), *moderate* (2), or *severe* (3).

The photos vary both in resolution and subjective quality, such as focus, lighting, and blur. In addition, as the photos can contain significant areas of background or areas that are otherwise irrelevant for diagnosis, we manually curated 962 of the original photos, generating 1748 image crops of representative diseased regions by visual inspection[4]. We used these crops to fine-tune an RoI detection network, and then bootstrapped our dataset by running this network on all images, extracting a further 2178 image crops. Both sets of image crops were then combined and paired with the labels for the 7 disease signs, resulting in a final dataset of 933 diagnoses from 285 patients, including 1237 original photos with corresponding 3926 image crops[5].

This final dataset was used to train our severity prediction network (Subsection 3.2). All crops were labelled with the overall diagnosis for the entire image, as we did not have labels for the individual crops. Despite this noisy labelling, the use of RoI detection and severity prediction in EczemaNet led to better performance than using the entire image (Subsection 4.2).

3 Method

Our EczemaNet pipeline consists of detecting RoI, making probabilistic predictions on all 7 disease signs over all crops simultaneously, and then combining these to predict the AD severity scores per image (Fig. 2). We made heavy use of transfer learning [30] to train on our medium-size dataset successfully: we fine-tuned both our RoI detection and severity prediction CNNs. The RoI detection

[4] RoI, of arbitrary size, were labelled by 3 volunteers given a set of 50 expert-labelled images, where 1 volunteer was instructed directly by an expert. 431 photos were deemed difficult to label by the volunteers and hence left out of our dataset.

[5] The full data pipeline is provided in Supp. Fig. 1.

was trained first, as otherwise it would not be able to provide relevant crops for the severity prediction network for end-to-end training. We used TensorFlow [1] for training and evaluation, starting with pretrained models in TensorFlow. Our code is available at https://github.com/Tanaka-Group/EczemaNet.

3.1 RoI (Region of Interest) Detection

Following the speed/memory/accuracy model selection guidelines from Huang et al. [15], we chose the Faster R-CNN model [18] to perform RoI detection for diseased areas.

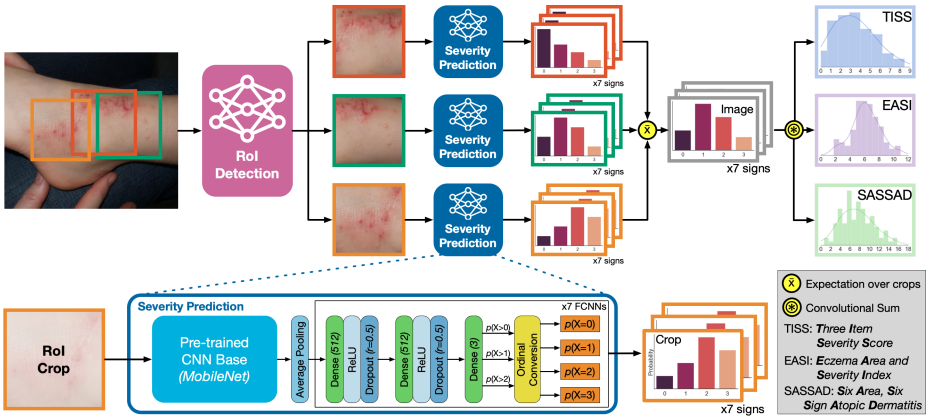


Fig. 2: EczemaNet overview. The RoI detection network extracts crops from an image. The severity prediction network makes probabilistic predictions for each disease sign in each crop. The averaged prediction over crops are then combined to form the final probabilistic prediction of the severity scores for the image.

3.2 Severity Prediction

Our severity prediction pipeline is composed of a pretrained CNN base and 7 fully-connected neural networks (FCNNs), each of which predicts the severity of one of the 7 disease signs. We reflect the ordinal nature of the labels by training the FCNNs with ordinal classification. The predicted severities are averaged over all crops to calculate a probabilistic distribution of the severity of each disease sign for the image. Finally, the predictions for the disease signs are combined to produce a probability distribution of the regional[6] severity scores (SASSAD, TISS and EASI) per image.

Here we describe characteristic features of EczemaNet in more detail.

Pretrained CNN base: Our base consists of all convolutional and pooling layers within MobileNet [14].

[6] In practice, EASI and SASSAD are assessed across different regions of the body, which we do not consider in this work.

Separate FCNNs: We use separate FCNNs per disease sign, as opposed to using one FCNN to predict all disease signs simultaneously.
Ordinal Classification: Instead of predicting the 4 severities independently for each sign as a 4-way classification, as is typically done, we model them using ordinal classification, which better reflects the ordinal nature of the severity. To predict the classes, X, for the diagnoses none ($X = 0$), mild ($X = 1$), moderate ($X = 2$) and severe ($X = 3$), we train 3 binary classifiers to output the probabilities, $p_0 = p(X > 0)$, $p_1 = p(X > 1)$ and $p_2 = p(X > 2)$. These probabilities are then converted into class probabilities for outcome X using a modification of Frank & Hall's method [8] with dependent classifiers [6]:
$p(X = 0) = 1 - p_0, \quad p(X = 1) = p_0(1 - p_1), \quad p(X = 2) = p_0 p_1(1 - p_2)$, and
$p(X = 3) = p_0 p_1 p_2$.
Expectation over Crops: We produce a single set of severity predictions for each disease sign over the entire image, by averaging the predictions over all crops[7]. Despite the high overlap between most crops, similarly to test-time data augmentation [3], we found that averaging over crops improved both accuracy and calibration (Subsection 4.2).
Multitask prediction: All 3 regional severity scores (TISS, EASI, SASSAD) are sums of subsets of the 7 disease signs (Fig. 1B). While it is possible to directly predict each of the regional severity scores, we treat prediction as a multitask problem, predicting the severity of all disease signs simultaneously, and then sum them[8] to calculate the final regional severity scores.

4 Experiments and Evaluation

Inference for a single image on CPU (Intel i9-9980HK) took 15.6s for the detection network and 1.6s for the severity prediction network. Our work is a proof-of-concept, and could feasibly run on a smartphone in a few seconds with, e.g., model compression techniques.

4.1 RoI (Region of Interest) Detection

We fine-tuned a pretrained Faster R-CNN model using the 962 manually curated original photos. With a train/validation/test ratio of 60:20:20, the manually curated photos were randomly split into 578:192:192 photos. It resulted in 1069:378:346 corresponding image crops, as each photo can contain a different number of image crops. The model was trained for 10^5 steps with a batch size of 1, using SGD with momentum = 0.9, with an initial learning rate of 3×10^{-4}, dropped to 3×10^{-5} after 90000 steps; no data augmentation was used. We weighted the localisation loss by a factor of 1.2, as our focus was to improve detection, rather than classification by Faster R-CNN, which was trained to detect the presence of AD.

[7] Crops were preprocessed by bilinearly resampling to 224×224px.
[8] We convolve the probability mass functions of the predicted severity of the 7 disease signs, assuming that the predictions are independent random variables.

We evaluated our model using the average precision (AP) score, the standard measure in object detection. The AP score measures the intersection between the ground truth and predicted boundaries, with a default overlap threshold of 50%. After tuning hyperparameters on our validation set, we tested our model using the test set of 192 images and obtained the AP score of 40.15%. We also performed a more qualitative evaluation to validate our trained model, and estimated that our model achieved a 10% false positive rate per image. We therefore concluded that our RoI detection network could generalise sufficiently well, and used it to extract more crops from the original data (Section 2).

4.2 Severity Prediction

We combined a pretrained MobileNet with 7 separate randomly initialised FC-NNs (for each disease sign), and trained all parameters to predict the severity of the 7 disease signs on the final pre-processed dataset, which contained 933 diagnoses from 285 patients, including 1237 original photos with 3926 corresponding image crops. We used 10-fold cross-validation with a 90:10 train/test split, stratified on patients, to train and assess severity prediction models. The models were trained for a maximum of 50 epochs (using early stopping) with a batch size of 32, using SGD with a learning rate of 1×10^{-4} and momentum $= 0.9$; no data augmentation was used. Dropout with $p = 0.5$ and a max ℓ_2-norm weight constraint with $c = 3$ were used to regularise all fully-connected layers [22]. To combat severe class imbalance, we weighted all prediction losses by the inverse of the empirical class probabilities.

We evaluated RMSE on EASI (the recommended severity score [20]) for EczemaNet (1.929 ± 0.019) and for its variations listed below to confirm the use of each characteristic aspect of our model design (Fig. 3A and Table 1).

Pretrained CNN Base: The choice of pretrained CNN base significantly impacts the performance of the prediction model. We evaluated a range of commonly used CNN architectures for the base: Inception-v3 [24], MobileNet [14], ResNet-50 [13], VGG-16, and VGG-19 [21]. Only EczemaNet with MobileNet consistently achieved an RMSE on EASI of < 2 including standard error.

Bootstrapped Dataset: Training EczemaNet with the 1748 manually labelled crops, plus the 2178 additional crops automatically extracted by our trained RoI detection network, achieved the lowest RMSE across all of our experimental conditions (1.929 ± 0.019), compared to 2.003 ± 0.024 when EczemaNet was trained with only the manually labelled crops.

Model architectures: We used a set of baselines (baseline and intercept-only) and ablations (listed in order of performance, Fig. 3A; Supp. Fig. 2):

EczemaNet	Our full model.
–Ordinal	4-way categorical classification *vs.* ordinal classification.
+Interaction	Sign interaction added by concatenating FCNN features *vs.* separate FCNN per sign.
–Separate FCNNs	A single FCNN for all 7 signs *vs.* separate FCNNs per sign.
–Crops	Using the entire image *vs.* averaging predictions over crops.
–Pretrained	Starting with random CNN weights *vs.* pretrained CNN weights.

Intercept-only	Predicting the average EASI in the training set.
Baseline	Predicting EASI from the whole image using regression.
–Multitask	Predicting EASI directly *vs.* summing predicted disease signs.

The full EczemaNet performs best, although some components have a lesser effect on the RMSE on EASI (Fig. 3A)[9]. In reverse order, multitask learning is the most important modelling choice, which possibly mitigates overfitting. The baseline model, which is a naive CNN-based approach, using regression on the whole image, performs almost the same as the intercept-only model, indicating the difficulty of our problem. Using pretrained weights and averaging over crops also play a large role in the good predictive performance of EczemaNet. Sharing FCNN parameters when modelling the 7 disease signs hurts performance slightly, perhaps due to interference between the 7 tasks. Finally, ordinal classification provides a small boost over categorical classification with MobileNet[14][10].

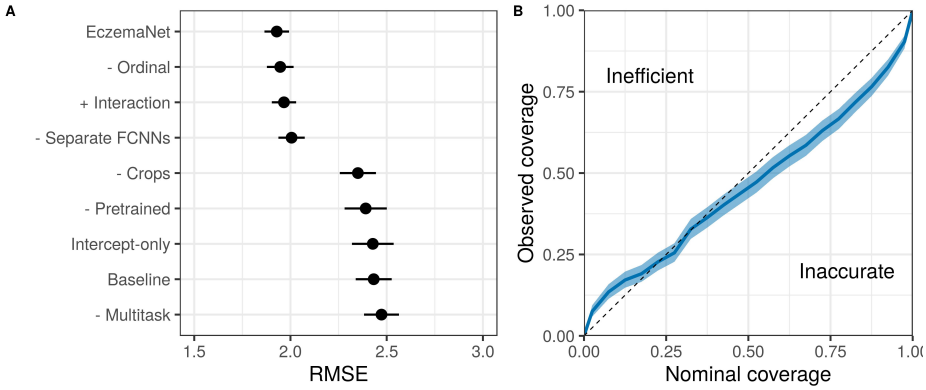


Fig. 3: A) RMSE (mean ± 1 standard error over cross-validation) on EASI across models. B) EASI calibration of highest density prediction intervals (coverage).

The coverage of EczemaNet (Fig. 3B) indicates well-calibrated prediction intervals for a NN [9]. The performance could be further improved by post-processing, such as quantile calibration, to make the predictive distribution sharper at the mode and with longer tails.

Achieving high accuracy on the regional severity scores is a major aim of our work for clinical relevance. It is also important to examine other metrics as well, particularly because of the class imbalance in the data. We calculated F_1 scores and Ranked Probability Scores (RPS) for all disease signs for all models that predict all 7 disease signs (Table 1). The F_1 score is the harmonic mean of precision and recall, and hence is less sensitive to class imbalance than recall.

[9] We also observed a similar ranking across models for SASSAD (Supp. Figure 3) and TISS (Supp. Figure 4), as well as across the individual signs.

[10] The selection of base architecture was determined experimentally. MobileNet[14] provided greater benefits over other base architectures including VGG-16/19 [21], ResNet-50 [12], and Inception-v3[23] (Supp. Fig. 5).

RPS is a strictly proper scoring rule corresponding to the MSE of the cumulative forecast distribution and the cumulative outcome distribution, and measures the calibration of ordinal forecasts. We observed approximately the same ranking of baselines/ablations as for RMSE on EASI, with no clear outliers, supporting our earlier assessment on their relative importance.

Table 1: F_1 score (top; ↑ is better) and RPS (bottom; ↓ is better) for models that predict all 7 disease signs. Mean ± 1 standard error over cross-validation.

Model	Cra.	Dry.	Ery.	Exc.	Exu.	Lic.	Oed.
Full	0.707 ± 0.013	0.443 ± 0.006	0.419 ± 0.004	0.480 ± 0.007	0.769 ± 0.008	0.404 ± 0.005	0.694 ± 0.007
–Pretrained	0.671 ± 0.013	0.242 ± 0.008	0.250 ± 0.009	0.269 ± 0.004	0.759 ± 0.008	0.234 ± 0.003	0.694 ± 0.007
–Separate FCNNs	0.696 ± 0.013	0.422 ± 0.006	0.405 ± 0.006	0.473 ± 0.005	0.768 ± 0.008	0.390 ± 0.005	0.690 ± 0.007
+Interaction	0.704 ± 0.013	0.454 ± 0.008	0.437 ± 0.005	0.491 ± 0.007	0.767 ± 0.008	0.388 ± 0.007	0.697 ± 0.007
–Ordinal	0.696 ± 0.013	0.453 ± 0.006	0.428 ± 0.007	0.470 ± 0.004	0.772 ± 0.008	0.404 ± 0.006	0.692 ± 0.008
–Crops	0.686 ± 0.012	0.369 ± 0.004	0.370 ± 0.006	0.289 ± 0.007	0.765 ± 0.007	0.317 ± 0.006	0.700 ± 0.007
Full	0.076 ± 0.003	0.136 ± 0.001	0.137 ± 0.001	0.128 ± 0.001	0.056 ± 0.002	0.151 ± 0.002	0.077 ± 0.002
–Pretrained	0.098 ± 0.003	0.164 ± 0.001	0.160 ± 0.002	0.178 ± 0.001	0.077 ± 0.002	0.181 ± 0.001	0.085 ± 0.002
–Separate FCNNs	0.080 ± 0.003	0.140 ± 0.001	0.142 ± 0.001	0.132 ± 0.001	0.057 ± 0.002	0.156 ± 0.002	0.079 ± 0.002
+Interaction	0.080 ± 0.003	0.141 ± 0.001	0.141 ± 0.001	0.131 ± 0.002	0.056 ± 0.002	0.156 ± 0.002	0.079 ± 0.002
–Ordinal	0.079 ± 0.003	0.139 ± 0.001	0.136 ± 0.001	0.130 ± 0.001	0.055 ± 0.002	0.149 ± 0.001	0.079 ± 0.002
–Crops	0.083 ± 0.004	0.154 ± 0.001	0.155 ± 0.002	0.163 ± 0.002	0.063 ± 0.002	0.165 ± 0.002	0.081 ± 0.002

5 Discussion

This paper presented EczemaNet, a CNN-based pipeline for evaluating eczema severity directly from camera images. EczemaNet consists of an RoI detection network, which extracts relevant crops from each image, and a severity prediction network, which predicts the severity of 7 disease signs for each crop. The probability distributions of severities are averaged over crops, and then combined to form a prediction of the 3 regional severity scores. EczemaNet achieved a low RMSE on EASI on a medium-size clinical dataset, and demonstrated well-calibrated prediction intervals. These results present a step towards standardising evaluation of objective AD severity scores for diverse dermatological research purposes, and could be applied to similar conditions, such as psoriasis.

Multiple sources of systematic errors, aside from random errors, can be considered to limit the performance of EczemaNet: mis-labelling due to inter- and intra-rater variability, discretisation error over a continuous outcome (severity), and errors arising from partial/noisy information (e.g., tactile diagnoses, out-of-focus images). While it is difficult to evaluate the effects of all of these sources of errors, Monte Carlo simulations of the measurement process suggested that rounding error alone could account for an RMSE of 0.6, which makes it unclear how much performance could be further improved on EczemaNet trained with our data. Future work could involve data augmentation while tackling the issue of the class imbalance.

A natural extension of the work presented here is to move beyond regional severity scores to predicting the overall severity scores. For a given area, EASI is the product of the intensity score (which we currently predict), and the area score. The area score could be predicted simultaneously with the intensity scores given the box labels identified by our RoI network. We encourage future clinical

trials to collect and share richer labels, such as pixel-level segmentations, to increase the breadth of tasks, such as segmentation, that can be automated using machine learning.

References

1. Abadi, M., Barham, P., Chen, J., Chen, Z., Davis, A., Dean, J., Devin, M., Ghemawat, S., Irving, G., Isard, M., et al.: Tensorflow: A system for large-scale machine learning. In: USENIX Symposium on Operating Systems Design and Implementation. pp. 265–283 (2016)
2. Alam, M.N., Munia, T.T.K., Tavakolian, K., Vasefi, F., MacKinnon, N., Fazel-Rezai, R.: Automatic detection and severity measurement of eczema using image processing. In: EMBC. pp. 1365–1368. IEEE (2016)
3. Ashukha, A., Lyzhov, A., Molchanov, D., Vetrov, D.: Pitfalls of in-domain uncertainty estimation and ensembling in deep learning. International Conference on Learning Representations (2020)
4. Berth-Jones, J.: Six area, six sign atopic dermatitis (sassad) severity score: a simple system for monitoring disease activity in atopic dermatitis. British Journal of Dermatology **135**, 25–30 (1996)
5. Brinker, T.J., Hekler, A., Enk, A.H., Klode, J., Hauschild, A., Berking, C., Schilling, B., Haferkamp, S., Schadendorf, D., Holland-Letz, T., et al.: Deep learning outperformed 136 of 157 dermatologists in a head-to-head dermoscopic melanoma image classification task. European Journal of Cancer **113**, 47–54 (2019)
6. Cardoso, J.S., Costa, J.F.: Learning to classify ordinal data: The data replication method. JMLR **8**(Jul), 1393–1429 (2007)
7. Esteva, A., Kuprel, B., Novoa, R.A., Ko, J., Swetter, S.M., Blau, H.M., Thrun, S.: Dermatologist-level classification of skin cancer with deep neural networks. Nature **542**(7639), 115–118 (2017)
8. Frank, E., Hall, M.: A simple approach to ordinal classification. In: ECML. pp. 145–156. Springer (2001)
9. Guo, C., Pleiss, G., Sun, Y., Weinberger, K.Q.: On calibration of modern neural networks. In: ICML. pp. 1321–1330 (2017)
10. Hameed, N., Shabut, A.M., Hossain, M.A.: Multi-class skin diseases classification using deep convolutional neural network and support vector machine. In: SKIMA. pp. 1–7. IEEE (2018)
11. Hanifin, J., Thurston, M., Omoto, M., Cherill, R., Tofte, S., Graeber, M., Evaluator Group, T.E.: The eczema area and severity index (easi): assessment of reliability in atopic dermatitis. Experimental Dermatology **10**(1), 11–18 (2001)
12. He, K., Zhang, X., Ren, S., Sun, J.: Deep residual learning for image recognition. In: 2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR). pp. 770–778 (2016)
13. He, K., Zhang, X., Ren, S., Sun, J.: Deep residual learning for image recognition. In: CVPR. pp. 770–778 (2016)
14. Howard, A.G., Zhu, M., Chen, B., Kalenichenko, D., Wang, W., Weyand, T., Andreetto, M., Adam, H.: Mobilenets: Efficient convolutional neural networks for mobile vision applications. arXiv:1704.04861 (2017)
15. Huang, J., Rathod, V., Sun, C., Zhu, M., Korattikara, A., Fathi, A., Fischer, I., Wojna, Z., Song, Y., Guadarrama, S., et al.: Speed/accuracy trade-offs for modern convolutional object detectors. In: CVPR. pp. 7310–7311 (2017)

16. Padilla, D., Yumang, A., Diaz, A.L., Inlong, G.: Differentiating atopic dermatitis and psoriasis chronic plaque using convolutional neural network mobilenet architecture. In: HNICEM. pp. 1–6 (2019)
17. Pal, A., Chaturvedi, A., Garain, U., Chandra, A., Chatterjee, R., Senapati, S.: Severity assessment of psoriatic plaques using deep cnn based ordinal classification. In: OR 2.0 Context-Aware Operating Theaters, Computer Assisted Robotic Endoscopy, Clinical Image-Based Procedures, and Skin Image Analysis, pp. 252–259. Springer (2018)
18. Ren, S., He, K., Girshick, R., Sun, J.: Faster r-cnn: Towards real-time object detection with region proposal networks. In: NeurIPS. pp. 91–99 (2015)
19. Schmitt, J., Langan, S., Deckert, S., Svensson, A., von Kobyletzki, L., Thomas, K., Spuls, P., et al.: Assessment of clinical signs of atopic dermatitis: a systematic review and recommendation. Journal of allergy and clinical immunology **132**(6), 1337–1347 (2013)
20. Schmitt, J., Spuls, P.I., Thomas, K.S., Simpson, E., Furue, M., Deckert, S., Dohil, M., Apfelbacher, C., Singh, J.A., Chalmers, J., et al.: The harmonising outcome measures for eczema (home) statement to assess clinical signs of atopic eczema in trials. Journal of allergy and clinical immunology **134**(4), 800–807 (2014)
21. Simonyan, K., Zisserman, A.: Very deep convolutional networks for large-scale image recognition. International Conference on Learning Representations (2015)
22. Srivastava, N., Hinton, G., Krizhevsky, A., Sutskever, I., Salakhutdinov, R.: Dropout: a simple way to prevent neural networks from overfitting. JMLR **15**(1), 1929–1958 (2014)
23. Szegedy, C., Vanhoucke, V., Ioffe, S., Shlens, J., Wojna, Z.: Rethinking the inception architecture for computer vision. In: 2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR). pp. 2818–2826 (2016)
24. Szegedy, C., Vanhoucke, V., Ioffe, S., Shlens, J., Wojna, Z.: Rethinking the inception architecture for computer vision. In: CVPR. pp. 2818–2826 (2016)
25. Thomas, K., Koller, K., Dean, T., o'Leary, C., Sach, T.H., Frost, A., Pallett, I., Crook, A., Meredith, S., Nunn, A., et al.: A multicentre randomised controlled trial and economic evaluation of ion-exchange water softeners for the treatment of eczema in children: the softened water eczema trial (swet). Health Technology Assessment **15**(8), 1–156 (2011)
26. Tschandl, P., Rosendahl, C., Kittler, H.: The ham10000 dataset, a large collection of multi-source dermatoscopic images of common pigmented skin lesions. Scientific data **5**, 180161 (2018)
27. Weidinger, S., Beck, L.A., Bieber, T., Kabashima, K., Irvine, A.D.: Atopic dermatitis. Nature Reviews Disease Primers **4**(1), 1 (2018)
28. Wolkerstorfer, A., De Waard van der Spek, F., Glazenburg, E., Mulder, P., Oranje, A.: Scoring the severity of atopic dermatitis: three item severity score as a rough system for daily practice and as a pre-screening tool for studies. Acta Dermato-Venereologica **79**, 356–359 (1999)
29. Wu, H., Yin, H., Chen, H., Sun, M., Liu, X., Yu, Y., Tang, Y., Long, H., Zhang, B., Zhang, J., et al.: A deep learning, image based approach for automated diagnosis for inflammatory skin diseases. Annals of Translational Medicine **8**(9) (2020)
30. Yosinski, J., Clune, J., Bengio, Y., Lipson, H.: How transferable are features in deep neural networks? In: NeurIPS. pp. 3320–3328 (2014)
31. Zuberbier, T., Orlow, S.J., Paller, A.S., Taïeb, A., Allen, R., Hernanz-Hermosa, J.M., Ocampo-Candiani, J., Cox, M., Langeraar, J., Simon, J.C.: Patient perspectives on the management of atopic dermatitis. Journal of Allergy and Clinical Immunology **118**(1), 226–232 (2006)