Interpreting Fine-Grained Dermatological Classification by Deep Learning

Sourav Mishra
The University of Tokyo
sourav@ay-lab.org

Hideaki Imaizumi
exMedio Inc.
imaq@exmed.io

Toshihiko Yamasaki
The University of Tokyo
yamasaki@ay-lab.org

Abstract

This paper analyzes a deep learning based classification process for common East Asian dermatological conditions. We have chosen ten common categories based on prevalence. With more than 85% accuracy in our experiments, we have tried to investigate why current models are yet to reach accuracy benchmarks seen in object identification tasks. Our current attempt sheds light on how deep learning based dermoscopic identification and dataset creation could be improved.

1. Introduction

Dermatological care is an established need in today's health scenario. With timely intervention, many problems can be resolved effectively. According to estimates by National Institutes of Health (NIH) in the US, one out of five US citizens are at a risk for developing a debilitating dermatological problem in their lifetimes [29]. If the skin anomalies are detected and treated early, the survival rate is close to 98%. Skin diseases such as contact dermatitis and ringworm, although not lethal, spread virulently [1, 5]. At a time when increasing demand for dermatological expertise is being observed, there is a constant under-supply in many countries. The number of dermatologists in the US has plateaued at about 3.6 doctors per 10,000 [12]. In Japan, telemedicine is being actively advocated [11, 7, 14].

In the absence of specialists, subjects commonly seek advice from general physicians. However, the diagnosis of a general physician is concurrent with a dermatologist only about 57% of the time, across the full spectrum of skin complaints [17]. There exists a large scope of error which could aggravate a subject's health situation.

Computer vision has been successful in many domains of health care. However, classical machine learning (ML) has not been effective in addressing the accurate identification of dermoscopic abnormalities. Rule based approaches were tedious to make computer aided detection a possibility. With the advent of deep learning, ML models are becoming increasingly better at such applications [13, 31]. Esteva *et al*. demonstrated dermatologist-level accuracy in detecting *Melanoma* [9]. However, their model was limited to detecting and grading only one type of condition. Similar studies have been published by Shrivastava *et al*. for *Psoriasis* [25]. Towards detecting multiple diseases, Park *et al*. tried to introduce crowd sourcing [20]. Currently, there is a scope to develop & confidently deploy deep learning based identification system for multiple disease categories, and reap its benefits by mobile based computing [23]. Although there have been some recent developments in classifying a spectrum of skin conditions, the results are quantitatively much lower than those in object detection models [18, 19, 15, 4]. Even with most of these projects employing transfer learning, the detection accuracy remains far below object recognition benchmarks. Hence, there is a scope to improve the outcomes by understanding how convolutional neural networks (CNN) interpret skin images during a fine-grained classification.

In this paper, we attempt to understand some factors for errors encountered during classification of dermoscopic images. Utilizing gradients-based guided back-propagation and class activation maps, we have tried to comprehend the classification mechanism by CNNs. In the process, we have developed an impression of why certain label pairs encounter higher miscategorizations. Our contribution to this topic is as follows:

- We investigate any differences in outcomes owing to different model sizes. Using existing state-of-the-art training methodologies, we compare different architectures trained on our dataset in terms of accuracy.
- We attempt a visual explanation of the classification errors in selected label pairs, by trying to identify the features that CNNs attribute to the wrong class.
- We attempt to identify the few important guidelines for better dataset creation from user submitted images.

The rest of this paper is organized as follows: We explain the process of dataset creation briefly in Section 2. Model learning with established state of the art training regimens

Label	Images
Acne	972
Alopecia	682
Blister	691
Crust	640
Erythema	689
Leukoderma	665
P. Macula	717
Tumor	790
Ulcer	782
Wheal	636
Total	7264

Table 1: Distribution of number of images across the ten most common skin complaints, matching observed outpatient statistics.

is covered in Section 3. Investigation of the model performance and case studies of irregular classification is covered in Section 4. We discuss recommendations regarding improvement of the classifier pipeline in Section 5 prior to concluding.

2. Dataset Preparation

For compiling a custom dermoscopic dataset of the East Asian skin type, we performed a systematic collection of images from volunteers. All the images were mostly bigger than 200 × 200 pixels and in JPEG format. Additionally, we sourced specimen images from medical centers. After anonymizing, these images were labeled by registered clinical practitioners. Images containing any identifying feature such as birthmark, tattoo, hospital-tags or indicative marks were excluded. With the advice of consultant physicians regarding prevalence, we filtered our choice to ten common dermoscopic labels. These are: (i) *Acne*, (ii) *Alopecia*, (iii) *Blister*, (iv) *Crust*, (v) *Erythema*, (vi) *Leukoderma*, (vii) *Pigmented Maculae*, (viii) *Pustule*, (ix) *Wheal* and (x) *Ulcer*. A total of 7264 images across the labels were chosen, and split randomly in the ratio of 5:1 for training and validation set. Table 1 offers further information on the quantitative distribution of these labels.

3. Model Learning

Before attempting to understand lacunae in the classification pipeline, our first step was to build a high-accuracy classifier with state of the art techniques. Our design was based on the PyTorch 1.0 framework, running on a single NVIDIA GPU (Tesla V100 16GB HBM2). We performed transfer learning with ResNet-34, ResNet-50, ResNet-101 and ResNet-152 pretrained on the ImageNet [22]. The batch size was set at 64 and binary cross entropy with logistic loss function was used.

Prior to model learning, we normalized the data with the recommended mean (0.485, 0.456, 0.406) and standard deviation (0.229, 0.224, 0.225). We performed dynamic in-memory augmentation by cropping, horizontal & vertical flips and zooming by appropriate transformations in the data loader. To get the best fit, we focused on optimizing the training process by tuning the learning rate α. In conventional methods, learning rate (LR) is chosen by user experience and allowed to decay monotonically during the training phase. However, we foresaw two problems: (1) A smaller than optimal learning rate could lead to stagnation of gradients early and (2) loss function could get stuck at a local minima midway. We addressed these problems by finding an optimal rate based on our data as well as using a mechanism of cosine rate annealing during training.

We implemented the LR range test based on Smith [26, 19], which used several mini-batches with increasing learning rates. The rate of change of loss was observed until it dropped dramatically and reach a point of inflexion. The learning rate α was chosen in the neighborhood of the optimal value for best performance. Figure 1 and 2 illustrate the LR range test performed on ResNet-152.

Following the determination of an optimal LR, we tried to train our network in two steps using stochastic gradient descent with restarts (SGD-R) [16]. In the first step, we froze the final layers of the chosen network architecture and pre-computed activations from our dataset prior to learning. This process can be perceived as a model conditioning step to fine tune the most active layer of the CNN. Using SGD-R, we performed cosine rate annealing for every epoch of training. The learning rate was reduced from our optimal value α_{opt} to near-zero, to again restart with α_{opt}. This modulation is governed by Equation 1,

$$\nu_t = \frac{1}{2}\left(1 + \nu cos\left(\frac{t\pi}{T}\right)\right) + \epsilon, \qquad (1)$$

where ν is the initial learning rate, t is the iteration over the epoch, T is the total number of iterations to cover a epoch, and ϵ is floating point error term. Up to 10 epochs were run using this method, until the validation accuracy stabilized. A schematic of this modulation is illustrated by Figure 3 for the model trained on ResNet-152.

In the second step, the models were unfrozen and the training was allowed to modify the whole network. Prior to commencing this step, we assigned different learning rates for different parts of the network. The initial one-third of the network, which captured rudimentary features from input images, was assigned a very low learning rate ($0.01\alpha_{opt}$). The final one-third, which observed a lot of volatility during training, was assigned the usual rate, α_{opt}. The midsection was assigned an intermediate value ($0.1\alpha_{opt}$). We also introduced cycle length multiplication of SGD-R, by a

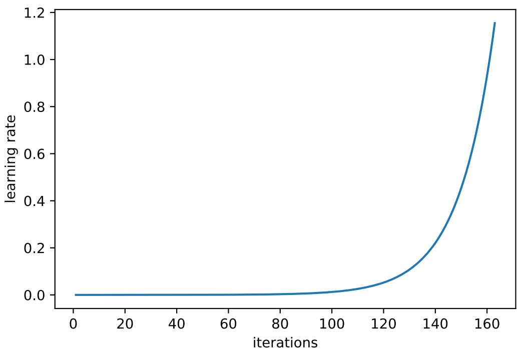


Figure 1: Learning rate α is systematically increased over several mini-batches and the losses determined. Concurrently, we also compute the rate of change of losses.

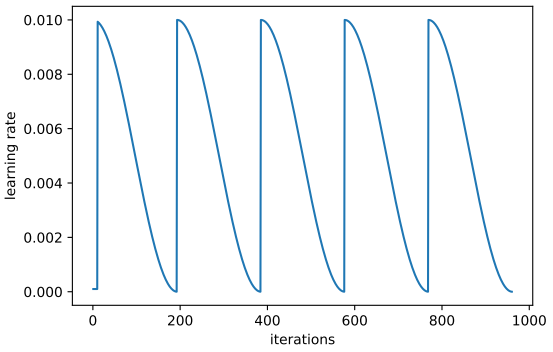


Figure 3: The architecture with all units frozen except the final layer, is trained with a variable learning rate. The rate decays following a cosine cycle until the end of the epoch, to again restart in the next one.

factor of two, also proposed in [16]. A schematic showing this modified learning rate schedule covering successively larger number of epochs, is given in Figure 4.

We can conceptualize the aforementioned step as follows: The bulk of model learning was done via SGD-R. With the model converging towards an optimum fit, disturbances were reduced by extending the LR cycle to cover several epochs [16]. This helped improving the quality of fit over closely similar labels. The periodic jumps minimized the scope of getting stuck at any local minima. Having differential learning rates reduced the chances for model to lose important pre-trained features. The results of training different ResNet architectures are shown in Table 2. A confusion matrix of the classification performance of ResNet-152 is illustrated in Figure 5.

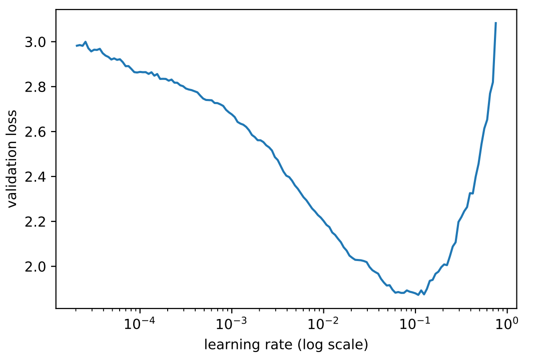


Figure 2: Plot of rate of change of losses. The optimal learning rate is chosen in the neighborhood of the point of inflexion, beyond which losses start increasing again.

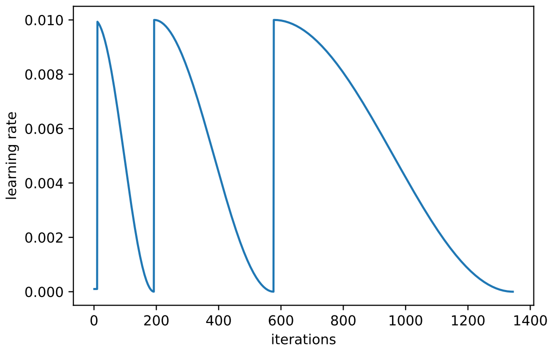


Figure 4: Rate annealing extends to cover progressively more number of epochs by cycle length multiplication (factor of 2).

Model	Peak Top-1 accuracy
ResNet-34	88.9%
ResNet-50	89.7%
ResNet-101	88.2%
ResNet-152	89.8%

Table 2: Results of optimal training of different architectures towards their best model fit. Different number of epochs & LR were employed to obtain a near consistent model accuracy in all the architectures.

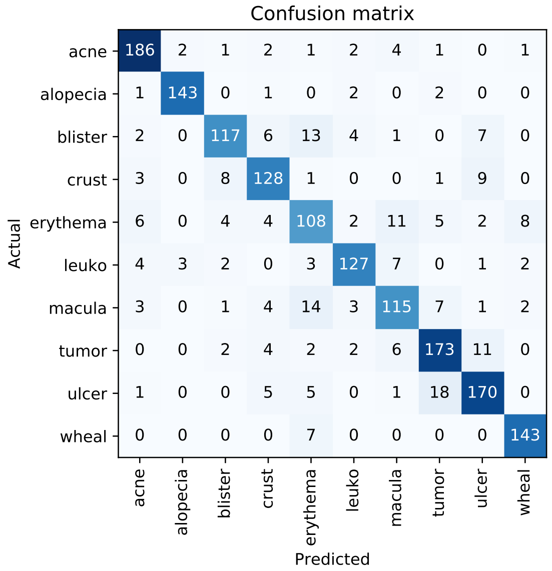


Figure 5: Confusion matrix of the classification in validation set, derived from the best performing architecture (ResNet-152).

4. Classification Case Studies

In Section 3, we trained several ResNet models to their best fits, to determine if there were any significant differences due to the architecture size. Using SGD-R & LR range test done over variable number of epochs, we found the validation accuracy to be converging to similar scores. Further proof of learning stability was seen in the loss curves (training and validation phases), which indicated limited possibility of further model fit. A plot representative of the ResNet-152 learning is shown in Figure 6. After pushing the model learning to their best fits, we could safely attribute the errors to the nature of images and incorrect annotations, if any.

To better understand the same, we focused on case studies for selected label pairs which tended to exhibit high miscategorizations in our experiments. Table 3 indicates prominent miscategorizatons derived from ResNet-152 model for example. We investigated these label pairs by using gradient-based class activation maps (Grad-CAM) [24] and guided backpropagation (GBP) [28]. In Figures 7-12 in the following sub-sections, the order of the images (L-R) are: sample, model prediction and true class label.

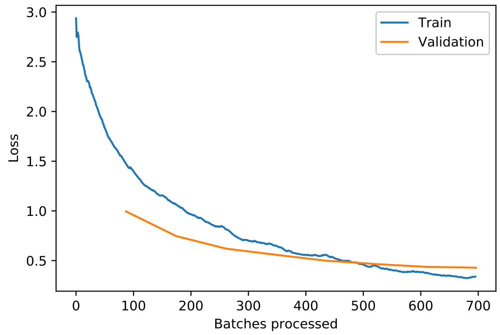


Figure 6: Both training and validation curves stabilize towards the end of the training. The flattened graphs indicate very less room for further drop in errors in the latter part of training.

Label 1	Label 2	Total
Ulcer	Tumor	29
P. Macula	Erythema	25
Blister	Erythema	17
Erythema	Wheal	15
Crust	Ulcer	14
Blister	Crust	14
P. Macula	Tumor	13
P. Macula	Leukoderma	10
Blister	Ulcer	07
Tumor	Erythema	07
Crust	Tumor	05

Table 3: The *Total* indicates the total number of incorrect predictions i.e., Label 1 predicted as 2, and vice-versa. The model is based on ResNet-152. Although different architectures and learning instances present different statistics, the trends exhibit a pattern in respect to the confused labels.

4.1. Ulcer & Tumor

Ulcer and *Tumor* have a high degree of prediction errors owing to similar planar attributes in their manifestation. In Case 1 (Figure 7), a sample of *Tumor* is presented which has been erroneously classified as *Ulcer*. There is a high degree of geometrical similarity between an average case of ulcer and this particular sample. The GBP plot delineates the circular lesion as an *Ulcer*. Our consultant physicians have had second thoughts by these predictions. It has been noted that the sample has a high degree of similarity with *Crust*, *Ulcer* and *Melanoma*. The spotting, although not detected in the class activation map, is similar to evolving secondary tumors seen in *Kaposi Sarcoma*. This is an example where the model could be highlighting the presence of labeling error or presence of a novel class.

A sample of benign *Ulcer* is presented in Case 2 (Figure 8) which has been misidentified as *Tumor*. It is interesting to note that the CNN identifies the inflammation around of lesion, to classify it as a *Tumor*. The region of

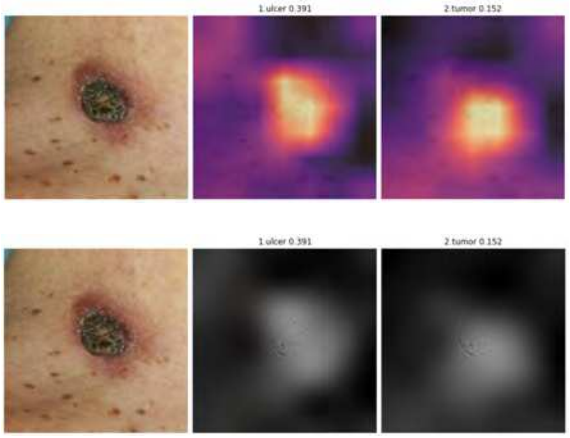


Figure 7: GradCAM (*top*) and GBP (*bottom*) plots of *Tumor* sample miscategorized as *Ulcer*.

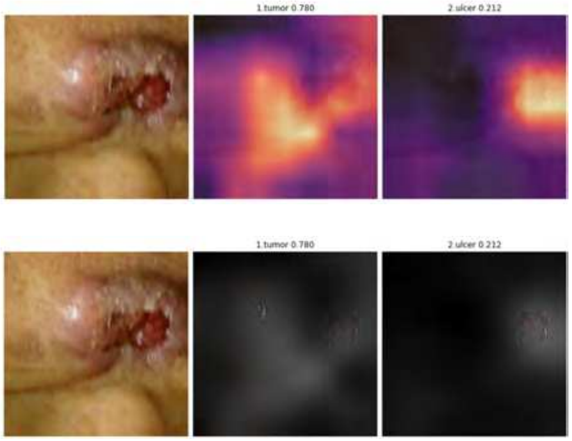


Figure 8: GradCAM (*top*) and GBP (*bottom*) plots of *Ulcer* sample miscategorized as *Tumor*.

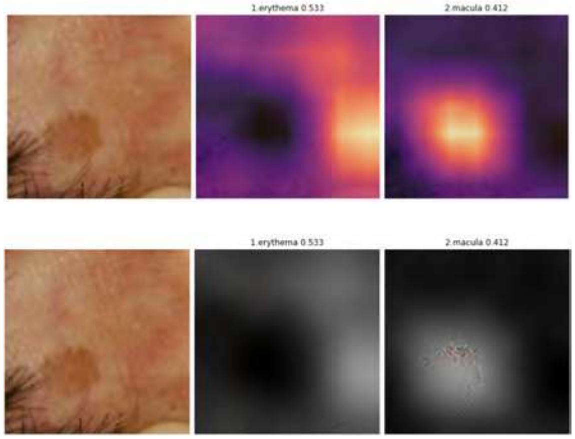


Figure 9: GradCAM (*top*) and GBP (*bottom*) plots of *P. Macula* sample miscategorized as *Erythema*.

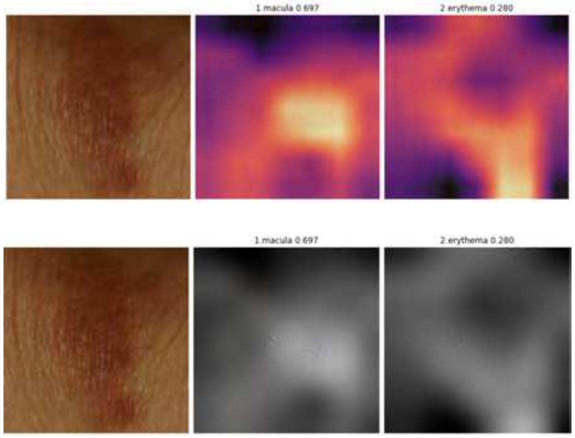


Figure 10: GradCAM (*top*) and GBP (*bottom*) plots of *Erythema* sample miscategorized as *P. Macula*.

interest (ROI) for ulcer is located with a certainty of 0.212. Lack of image centering, compounded by illumination artifact and presence of inflammation ranks the *Tumor* label much higher.

4.2. Pigmented Macula & Erythema

Pigmented Macula and *Erythema* have some of the highest rates of confusion, across our learned models. To analyze classification performance, we exhibit Cases 3 & 4, where the first one has true label of *P. Macula* and the second belongs to *Erythema*. In Case 3 (Figure 9), although the *P. Macula* has been detected with a probability of 41%, the larger pigmentation patch neighboring the spot leads to the model predicting the sample as *Erythema*. The GradCAM and GBP plots highlight the region of brown hypopigmentation coherently. In Case 4 (Figure 10), although there are no other visual features except the Erythemic pigmentation, the shape of the lesion exerts a strong bias on miscategorization as *P. Macula*. We hypothesize that presence of pigmentation in the field of view (FOV) is a strong factor in miscategorization of these labels.

4.3. Ulcer & Crust

The labels of *Ulcer* and *Crust* present an interesting challenge in classification. These labels are visually close. Often *Crust* appears during the healing process of *Ulcers*. Hence there is a strong visual correlation between the two. In the absence of chronological history of diagnosis, it is possible even for human interpreters to fail in categorizing these cases accurately. Case 5 (Figure 11) shows such an example. In the absence of treatment information, it may not be straightforward to ascertain whether the lesion is a *Ulcer* in recessing phase, or a existing *Crust*. Case 6 (Figure 12) from a lesion such as *Vesicle* or *Bulla*, has been

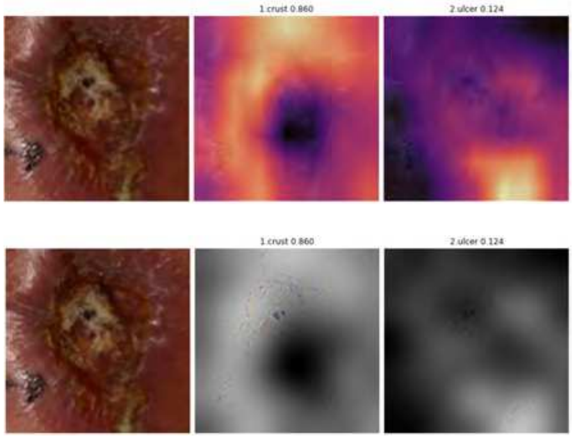


Figure 11: GradCAM (*top*) and GBP (*bottom*) plots of *Ulcer* sample miscategorized as *Crust*.

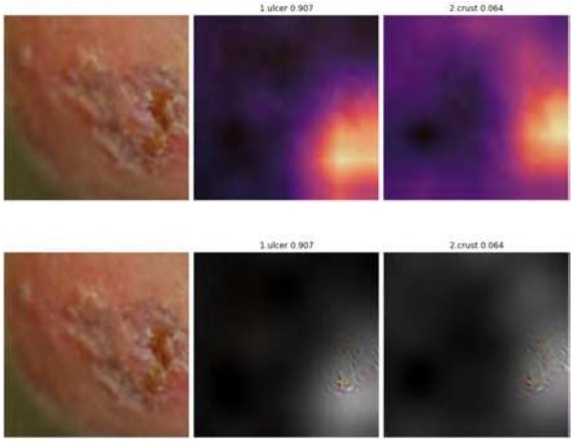


Figure 12: GradCAM (*top*) and GBP (*bottom*) plots of *Crust* sample miscategorized as *Ulcer*.

miscategorized as an Ulcer. The image in question is quite close to the *Ulcer* label visually; The GradCAM plots recognize the crater of the *Crust* as a identifying factor towards the wrong label. The predictions could have improved if the lesion was centered in the field of view.

5. Discussion

In Section 3, we empirically showed most available ResNet architectures can be trained to perform robustly by tuning hyperparameter settings. In Section 4, we showed cases of irregularities due to the nature of fine-grained differences in the data. These may not be easily remediated in the near future. ML models are not robust to accurately identify multiple lesions in a wide field of view. They can perform poorly in the presence of image blur, low light and noise. This is important from the end user standpoint where captured images could be very different to those taken in standardized experimental conditions.

There are some caveats to our experimental design as well. We have focused primarily on East Asian race. The model learning may not suitably generalize to other racial types. We designed our classifier on the ten most prevalent dermoscopic labels. Our model is not reliable towards other disease categories. Even if we managed to train on all available dermoscopic labels, the only tell-tale indicators to the presence of a novel category could be a higher discrepancy in prediction than usual, in which case a specialist opinion cannot be excluded (*sic.* Figure 7). We propose a few essential steps towards curating a dermoscopic dataset. From our experience, they have proved valuable to making models more resilient to misdiagnosis.

5.1. Balanced Training Set

The most significant contributor to model performance, and consequently any automation thereafter, is the nature of sample distribution in the training set. A balanced training set exhibits a true macro-average. This factor gains prominence in dataset creation from user submitted images. Even if classes are created with proportional uniformity, the database needs to be periodically checked for inter-class balance and duplicate information. To illustrate our point, consider Table 4. These statistics are generated by training models on an unbalanced distribution of voluntary submissions, closely mimicking recent out-patient statistic. The high values of accuracy across all the architectures still convey a false notion, since the metric is strongly dependent on the dominant class, i.e. *Erythema*. A confusion matrix in Figure 13, illustrates this point more clearly.

To balance our dataset, we collated user-submitted images with clinical samples and performed simple data augmentation (ref. Sec. 2), until our image classes were comparable to each other in size and diversity. If new pieces of user-submitted data are added exclusively to training set, periodic reshuffling is recommended. Image augmentation can be built into the ML model as on-the-fly transformations instead of maintaining a separate set of images. Although skin lesions can be modelled to look very realistic [2, 3, 21], we avoided using them since expert opinion is still divided on their use in real world applications [10].

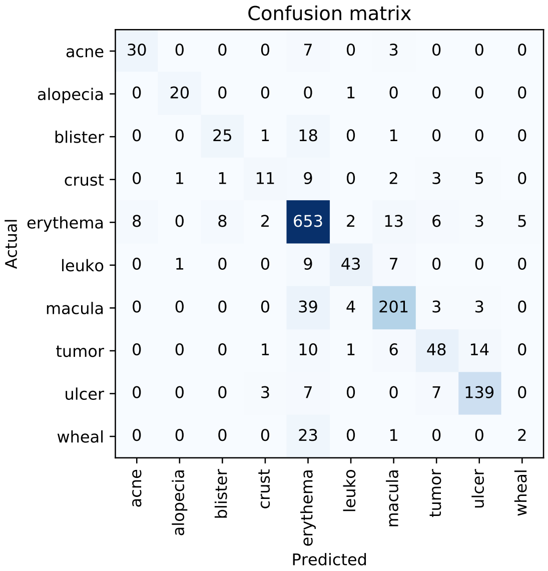
5.2. Optimizing Field of View

The field of view can affect the quality of detection significantly. Although, it is not unusual for dermoscopic images to present more than one lesion in the ROI, reducing the FOV and object distance can allow us to identity the lesion of interest and get better predictions. According to Sprawls [27], FOV reduction can even counteract some of the effects of blur and noise in medical images for machine learning.

In Figure 14, we illustrate the effect of FOV reduction

Model	Peak Top-1 accuracy
ResNet-34	83.40%
ResNet-50	83.26%
ResNet-101	83.19%
ResNet-152	84.90%

Table 4: Macro-average performance of different model architectures from user submitted images only described in Sec. 5.1. The high overall accuracy overshadows any poor class prediction, by a significant biasing.



Actual \ Predicted	acne	alopecia	blister	crust	erythema	leuko	macula	tumor	ulcer	wheal
acne	30	0	0	0	7	0	3	0	0	0
alopecia	0	20	0	0	0	1	0	0	0	0
blister	0	0	25	1	18	0	1	0	0	0
crust	0	1	1	11	9	0	2	3	5	0
erythema	8	0	8	2	653	2	13	6	3	5
leuko	0	1	0	0	9	43	7	0	0	0
macula	0	0	0	0	39	4	201	3	3	0
tumor	0	0	0	1	10	1	6	48	14	0
ulcer	0	0	0	3	7	0	0	7	139	0
wheal	0	0	0	0	23	0	1	0	0	2

Figure 13: Confusion matrix of the multi-class classification, derived from an unbalanced dataset of user submitted images only described in Sec. 5.1. ResNet-152 architecture was utilized.

in capturing the same lesion from two different object distances. The model prediction is unambiguous when the lesion is centered and at a shorter distance, with all other environment factors remaining the same. Since most of the user submitted images are captured via camera phone, application developers can be encouraged to add features to help users capture an optimal FOV. Image datasets require more intensive efforts by cropping the main ROI by trained clinicians. Crowd-sourcing can also be an effective means to be able to locate the correct regions, as shown for medical images in [6].

5.3. Illumination & Gamma Correction

Illumination artifacts are present in images which have not been captured under uniform lighting. These are common in user submitted images as bright regions in the image space or a high reflectance patch. The sources for such errors can be attributed insufficient lighting, presence of shadow regions or a camera flash. Because they are present in many labels, they can be sources of spurious prediction. Our proposal to alleviate specular reflection is to capture two set of images: one in ambient conditions and the other in presence of camera flash. The two images can be processed jointly as proposed in DiCarlo *et al.* [8], to produce a surface reflectance map. This meta-information can alleviate some instances of the artifacts. If several images in a database are known to have been captured in the same environment, lighting-independent models of illumination can be reconstructed by Bi-directional reflectance distribution function (BRDF) as proposed by Yu *et al.* [30]. Discussions about this scheme is beyond the scope of our current paper. Since CNNs capture geometrical features during the model fit, gamma adjustment ($\gamma = 1.2 - 1.5$) has been seen to create better instances of predictions. Figure 15 shows an example of *tumor* sample where prediction was significantly improved by gamma balancing. Shapes and contour detection is a factor towards better model performance, as seen in Section 4.

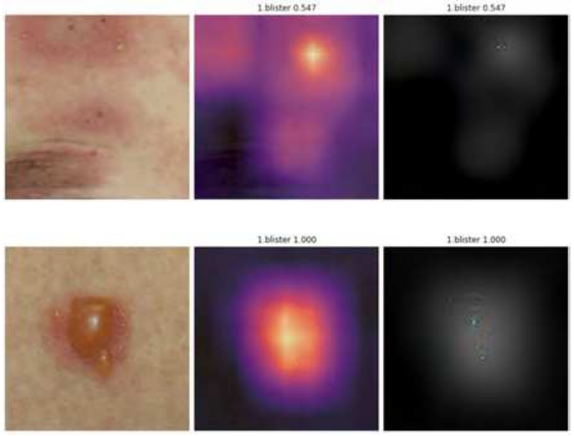


Figure 14: Large (*top*) and small (*bottom*) FOV of a *Blister* sample can have a significant difference in the quality of prediction (Ref. Sec. 5.2). The *Blister* was unambiguously predicted when the object distance was reduced.

6. Conclusion

This paper elucidates that several common skin problems can be successfully detected with deep learning techniques. We have shown that performance can be made model agnostic by adopting efficient hyperparameter settings. We have attempted to explain the nature of discrepancies in many label pairs and ways to alleviate them in user submitted images. We have commented on several best practices from our experience handling diagnos-

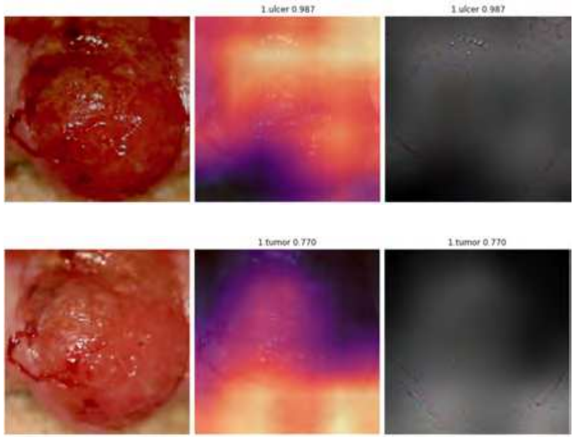


Figure 15: Original (*top*) and gamma corrected (*bottom*) images of a tumor, as described in Sec. 5.3. The sample prediction was corrected after gamma balancing.

tic automation. We hope that these prove useful as guidelines when creating new dermoscopic datasets in future. Our repository of user submitted images is available at https://github.com/souravmishra/ISIC-CVPRW19.

The model learning on dermatological images is only skin deep, if not complemented by information on patient history and other existing symptoms. Several factors are considered when a specialist makes a diagnosis. Confidence on ML-based diagnostics is precluded until more holistic information about the subject is available to analyze. Although not meant to replace a specialist, ML-based methods can prove to be effective diagnostic aids in screening patients.

References

[1] Oma N Agbai, Kesha Buster, Miguel Sanchez, Claudia Hernandez, Roopal V Kundu, Melvin Chiu, Wendy E Roberts, Zoe D Draelos, Reva Bhushan, Susan C Taylor, et al. Skin cancer and photoprotection in people of color: a review and recommendations for physicians and the public. *Journal of the American Academy of Dermatology*, 70(4):748–762, 2014.

[2] Christoph Baur, Shadi Albarqouni, and Nassir Navab. Generating highly realistic images of skin lesions with gans. In *OR 2.0 Context-Aware Operating Theaters, Computer Assisted Robotic Endoscopy, Clinical Image-Based Procedures, and Skin Image Analysis*, pages 260–267. Springer, 2018.

[3] Alceu Bissoto, Fábio Perez, Eduardo Valle, and Sandra Avila. Skin lesion synthesis with generative adversarial networks. In *OR 2.0 Context-Aware Operating Theaters, Computer Assisted Robotic Endoscopy, Clinical Image-Based Procedures, and Skin Image Analysis*, pages 294–302. Springer, 2018.

[4] Noel CF Codella, Q-B Nguyen, Sharath Pankanti, DA Gutman, Brian Helba, AC Halpern, and John R Smith. Deep learning ensembles for melanoma recognition in dermoscopy images. *IBM Journal of Research and Development*, 61(4/5):5–1, 2017.

[5] Sean M Dawes, Sheena Tsai, Haley Gittleman, Jill S Barnholtz-Sloan, and Jeremy S Bordeaux. Racial disparities in melanoma survival. *Journal of the American Academy of Dermatology*, 75(5):983–991, 2016.

[6] A Garcıa Seco de Herrera, Antonio Foncubierta-Rodríguez, Dimitrios Markonis, Roger Schaer, and Henning Müller. Crowdsourcing for medical image classification. In *Annual congress SGMI*, volume 2014, 2014.

[7] Itaru Dekio, Eisuke Hanada, Yuko Chinuki, Tatsuya Akaki, Mitsuhiro Kitani, Yuko Shiraishi, Sakae Kaneko, Minao Furumura, and Eishin Morita. Usefulness and economic evaluation of adsl-based live interactive teledermatology in areas with shortage of dermatologists. *International journal of dermatology*, 49(11):1272–1275, 2010.

[8] Jeffrey M DiCarlo, Feng Xiao, and Brian A Wandell. Illuminating illumination. In *Color and Imaging Conference*, volume 2001, pages 27–34. Society for Imaging Science and Technology, 2001.

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

[10] Jan Philip Göpfert, Heiko Wersing, and Barbara Hammer. Adversarial attacks hidden in plain sight. *arXiv preprint arXiv:1902.09286*, 2019.

[11] Hideaki Imaizumi, Akio Watanabe, Hiromi Hirano, Masatoshi Takemura, Hideyuki Kashiwagi, and Shinichiro Monobe. Hippocra: Doctor-to-doctor teledermatology consultation service towards future ai-based diagnosis system in japan. In *2017 IEEE International Conference on Consumer Electronics-Taiwan (ICCE-TW)*, pages 51–52. IEEE, 2017.

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

[13] Alex Krizhevsky, Ilya Sutskever, and G Hinton. Imagenet classification with deep convolutional networks. In *Proceedings of the Conference Neural Information Processing Systems (NIPS)*, pages 1097–1105, 2012.

[14] Rosilene Canzi Lanzini, Robyn S Fallen, Judy Wismer, and Hermenio C Lima. Impact of the number of dermatologists on dermatology biomedical research: a canadian study. *Journal of cutaneous medicine and surgery*, 16(3):174–179, 2012.

[15] Haofu Liao. A deep learning approach to universal skin disease classification. *University of Rochester Department of Computer Science, CSC*, 2016.

[16] Ilya Loshchilov and Frank Hutter. Sgdr: Stochastic gradient descent with warm restarts. In *ICLR 2017*, 2017.

[17] Brooke A Lowell, Catherine W Froelich, Daniel G Federman, and Robert S Kirsner. Dermatology in primary care: prevalence and patient disposition. *Journal of the American Academy of Dermatology*, 45(2):250–255, 2001.

[18] Sourav Mishra, Toshihiko Yamasaki, and Hideaki Imaizumi. Supervised classification of dermatological diseases by deep learning. *arXiv preprint arXiv:1802.03752*, 2018.

[19] Sourav Mishra, Toshihiko Yamasaki, and Hideaki Imaizumi. Improving image classifiers for small datasets by learning rate adaptations. *arXiv preprint arXiv:1903.10726*, 2019.

[20] Andrew J Park, Justin M Ko, and Robert A Swerlick. Crowdsourcing dermatology: Dataderm, big data analytics, and machine learning technology. *Journal of the American Academy of Dermatology*, 78(3):643–644, 2018.

[21] Fábio Perez, Cristina Vasconcelos, Sandra Avila, and Eduardo Valle. Data augmentation for skin lesion analysis. In *OR 2.0 Context-Aware Operating Theaters, Computer Assisted Robotic Endoscopy, Clinical Image-Based Procedures, and Skin Image Analysis*, pages 303–311. Springer, 2018.

[22] Olga Russakovsky, Jia Deng, Hao Su, Jonathan Krause, Sanjeev Satheesh, Sean Ma, Zhiheng Huang, Andrej Karpathy, Aditya Khosla, Michael Bernstein, et al. Imagenet large scale visual recognition challenge. *International journal of computer vision*, 115(3):211–252, 2015.

[23] Pranjal Sahu, Dantong Yu, and Hong Qin. Apply lightweight deep learning on internet of things for low-cost and easy-to-access skin cancer detection. In *Medical Imaging 2018: Imaging Informatics for Healthcare, Research, and Applications*, volume 10579, page 1057912. International Society for Optics and Photonics, 2018.

[24] Ramprasaath R Selvaraju, Michael Cogswell, Abhishek Das, Ramakrishna Vedantam, Devi Parikh, and Dhruv Batra. Grad-cam: Visual explanations from deep networks via gradient-based localization. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 618–626, 2017.

[25] Vimal K Shrivastava, Narendra D Londhe, Rajendra S Sonawane, and Jasjit S Suri. Reliable and accurate psoriasis disease classification in dermatology images using comprehensive feature space in machine learning paradigm. *Expert Systems with Applications*, 42(15-16):6184–6195, 2015.

[26] Leslie N. Smith. Cyclical learning rates for training neural networks. *2017 IEEE Winter Conference on Applications of Computer Vision (WACV)*, pages 464–472, 2017.

[27] P Sprawls. Optimizing medical image contrast, detail and noise in the digital era. *Medical Physics International*, 2(1), 2014.

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

[29] Robert S Stern. Prevalence of a history of skin cancer in 2007: Results of an incidence-based model. *Archives of dermatology*, 146(3):279–282, 2010.

[30] Yizhou Yu, Paul Debevec, Jitendra Malik, and Tim Hawkins. Inverse global illumination: Recovering reflectance models of real scenes from photographs. In *Siggraph*, volume 99, pages 215–224, 1999.

[31] Ning Zhang, Jeff Donahue, Ross Girshick, and Trevor Darrell. Part-based r-cnns for fine-grained category detection. In *European conference on computer vision*, pages 834–849. Springer, 2014.