Robustness of Deep Learning Models in Dermatological Evaluation: A Critical Assessment

Sourav MISHRA††, Subhajit CHAUDHURY†, Hideaki IMAIZUMI†††, Nonmembers, and Toshihiko YAMASAKI†, Member

SUMMARY Our paper attempts to critically assess the robustness of deep learning methods in dermatological evaluation. Although deep learning is being increasingly sought as a means to improve dermatological diagnostics, the performance of models and methods have been rarely investigated beyond studies done under ideal settings. We aim to look beyond results obtained on curated and ideal data corpus, by investigating resilience and performance on user-submitted data. Assessing via few imitated conditions, we have found the overall accuracy to drop and individual predictions change significantly in many cases despite of robust training.

key words: deep learning, dermatology, interpretability, robustness

1. Introduction

A growing need in our current healthcare scenario is timely dermatological attention. This need has been due to increasing incidences of skin diseases and an observed under-supply of specialists. Approximately 1.9 billion people worldwide are having skin condition, with such diseases being the fourth-most common source of morbidity [1]. In the US, there are only 3.6 dermatologists per 10,000 people [2]. Japan is actively advocating telemedicine as a remedial measure to bridge shortages in remote areas [3]–[5].

The spectrum of diseases in dermatology is sufficiently wide. With timely intervention, however, many problems can be resolved effectively. According to surveys by National Institutes of Health (NIH) in the US, one in every five American is at a risk for developing a debilitating skin problem. If the detection is made early, the survival rate is close to 98% [6]. At a time when the requirements are increasing, a constant under-supply is leading to long waiting queues [7]. In the absence of immediate attention, patients tend to consult general practitioners. Their diagnostic accuracy is between 24–70% and concurrency with specialist opinions around 57% [1], [8]. Even with specialists, opinions could differ at various stages. Hence, there exists a large scope for error.

Computer vision has been very successful recently in many domains due to the advent of deep learning [9]. They have circumvented classical rule-based approaches, which made identification tedious. Deep learning has demonstrated success in this current domain as well. Esteva et al. demonstrated dermatologist-level accuracy in detecting Melanoma [10]. Their model was, however, limited to grading only one type of condition. Similar studies have been conducted by Shrivastava et al. in detecting Psoriasis [11]. In addressing several diseases, Park et al. used the crowd-sourcing model for better anomaly prediction [12]. Liu et al. have managed to achieve about 70% accuracy on 26 diseases with a differential diagnosis system [1]. There is a need for robust expert systems to assist clinicians. Recent forays into building such systems are based on curated data [13]–[15]. We believe hand-picked multimedia obtained under ideal conditions are not true representation of samples seen in the clinical workflow.

In this paper, we attempt to evaluate the robustness of classifiers for ten common skin conditions seen in the East Asian skin type. We describe optimal training strategies, first verified against a standard computer vision dataset. Thereafter, we attempt to interpret the model performance by attention mechanisms, and reasons for failure by recreating artifacts and careful analysis. The contributions of this paper are as follows:

• We have demonstrated method(s) to rapidly train deep learning systems to high accuracy for detecting skin conditions (from a single racial type). We have investigated the role of model size towards overall accuracy.
• We have attempted to understand the dermoscopic classifier’s decision on wrong predictions via attention-based mechanisms.
• We have investigated the performance under the presence of common artifacts such as impulse noise and soft focus (blur).
• We have attempted to quantify the effects of distribution shift on our learned model.
• We have ascertained tendency of models to get biased by employing tests designed on synthetic data.

2. Data and Methods

2.1 Skin Data

A systematic collection of dermatological images was made with the consent and cooperation of volunteers belonging to the East Asian race. The images had different dimensions and sizes, but mostly larger than $200 \times 200$ and in
Table 1 Sample distribution summary

| Label   | Samples |
|---------|---------|
| Acne    | 971     |
| Alopecia| 681     |
| Blister | 690     |
| Crust   | 639     |
| Erythema| 689     |
| Leukoderma | 664 |
| P. Macula| 717     |
| Tumor   | 790     |
| Ulcer   | 782     |
| Wheal   | 636     |
| Total   | 7259    |

Table 2 SD-198 grouped for distribution shift study.

| Label          | Samples |
|----------------|---------|
| Acne           | Acne Keloidalis Nuchae, Acne Vulgaris, Steroid Acne, Favre Racouchot, Nevus Comedonicus, Pomade Acne |
| Alopecia       | Alopecia Areata, Androgenetic Alopecia, Follicular Mucinosis, Kerion, Scar Alopecia. |
| Blister        | Dyshidrosiform Eczema, Hailey Disease, Herpes Simplex, Herpes Zoster, Varicella, Mucha Habermann disease |
| Crust          | Angular Cheilitis, Bowen’s Disease, Impetigo |
| Erythema       | Acute Eczema, Candidiasis, Erythema Ab Igne, Ery. Anulare Centrifugum, Ery. Craque, Ery. Multiforme, Rosacea, Exfoliative Erythoderma |
| Leukoderma     | Balanitis Xerotica Obl., Beau’s Lines, Halo Nevus, Leukonychia, Pityriasis Alba, Vitiligo |
| P. Macula      | Actinic Solar Damage, Becker’s Nevus, Blue Nevus, Cafe Au Lait Macula, Compound Nevus, Congenital Nevus Dermatosis Nigra, Epidermal Nevus, Green Nail |
| Tumor          | Angioma, Apocrine Hydrocystoma, Lipoma, Dermatofibroma, Digital Fibroma, Fibroma, Leiomyoma |
| Ulcer          | Aphthous Ulcer, Behcet’s Disease, Ulcer, Stasis Ulcer, Mal Perforans, Pyoderma Gangrenosum, Syringoma |
| Wheal          | Urticaria, Stasis Edema |

JPEG format. Additional samples were sourced from medical centers and affiliated contributors within agreed frameworks of data reuse. These samples were labeled without any modifications. Any photograph containing identifying features (such as face, birthmark, tattoo, hospital-tags, etc.) was excluded. With statistics regarding local prevalence, we filtered our choice to ten common labels. These are: (i) Acne, (ii) Alopecia, (iii) Blister, (iv) Crust, (v) Erythema, (vi) Leukoderma, (vii) Pigmented Maculae, (viii) Tumor, (ix) Ulcer and (x) Wheal. Table 1 gives an overview of the distribution of classes in our study.

To examine the effect of distribution shift, we chose the SD-198 dataset [16]. It was ideal for our study since it was also composed of user submitted images. Since a one-to-one correspondence did not exist between classes in this dataset and our collected images, a dermatologist helped group relevant labels to our experimental design. Hundred samples were selected at random from these composite new classes and tested with our model. Information about this grouping is shown in Table 2.

2.2 Method Validation

In our study, the number of samples available were fewer compared to standardized computer vision datasets. Smaller data corpus could lead models to exhibit bias or overfitting. The absence of benchmarks made evaluating training methodology difficult. In healthcare domain, we anticipated rapid re-training as a requirement too. Hence, learning paradigms were needed that demonstrated reasonable accuracy, speed and repeatable performance. To cover these requirements, we chose to validate the classifier design on CIFAR-10 vision dataset [17] initially and we adopt the best practices thereafter.

We used a vanilla multi-class classification model as a baseline. The models were built on PyTorch with commonly recommended best practices [18]. Pretrained ResNet-34, ResNet-50, ResNet-101, ResNet-152, DenseNet-121, ResNext-50 (32 × 4d) and ResNext-101 (32 × 8d) were chosen as candidate starting points [19]. We chose serialized architectures over neural architecture search (NAS) based models since it was easy for us to design layer wise improvements in our learning scheme. Designing improvements on search-based architectures such as NASNet and AutoML is beyond the scope of the current study. A single GPU (NVIDIA Volta™ V100 16GB HBM2) was used. We normalized data with the recommended mean (0.4914, 0.4822, 0.4465) and standard deviation (0.2023, 0.1994, 0.2010). We chose a more traditional data split of 5:1 before training, keeping the dataset size in mind. We performed dynamic in-memory augmentation such as crop, random zoom, horizontal & vertical flips in the data-loader. All the models were trained to at least 90% validation accuracy for a fair comparison to the improved. The training was commenced with a learning rate of \( \alpha = 0.01 \) and restarted manually with lower value (\( \alpha = 0.001 \)), whenever early stopping halted learning. The accuracy and training time serving as baseline are elucidated in Table 3.

2.3 Improved Training Scheme

We optimized learning performance by exploiting the properties of the learning rate \( \alpha \). Conventional wisdom dictates monotonic decrease of learning rate(s) (LR). This straightforward approach, however, suffers from vanishing or exploding gradients often. We introduced an initial rate finder inspired by Smith et al. [23]. Our implementation of range test used several random mini-batches with linearly increas-
Initial learning rate was found by systematically increasing $\alpha$ over an interval and tracking the second moment of loss function. Tracking the second moment of the loss function, we could identify a good initial LR, as seen in Fig. 1 [22].

We carried out the training in two steps using stochastic gradient descent with warm restarts (SGD-R) [24]. As a first step, we froze the network except the final fully-connected (FC) layer. This step was a coarse conditioning step over about 10 epochs to fine-tune the most active layer. Using SGD-R, we performed cosine rate annealing for every epoch of training. Each time, the learning rate $\alpha_{opt}$ was reduced from its selected value to near-zero, to again restart with $\alpha_{opt}$ at the commencement of the next epoch. The modulation is governed by Eq. (1):

$$v_t = \frac{1}{2} \left(1 + \cos \left(\frac{\pi t}{T}\right) \right) + v_{\text{min}}, \quad (1)$$

where $v_t$ is the learning rate at the current iteration, $v$ is the initial learning rate, $t$ is a iteration in the epoch, and $T$ is the total number of iterations covering an epoch. In the subsequent step, the model was unfrozen to allow changes to all the parameters. To avoid significant perturbations, discriminative learning rates (DLR) were assigned to different sections of the network [25]. The initial one-third of the network, which captured rudimentary features from the input, was assigned lowest LR values ($\alpha = 0.01 \alpha_{opt}$) with LR gradually increasing to higher values towards higher layers. We also modified the rate schedule to involve cycle length multiplication (CLM) by a factor of 2. With the model converging towards its optimum fit, disturbances were reduced by extending the LR cycle to cover several epochs. The periodic jumps minimized the scope of getting stuck at saddle points. Having discriminative learning rates reduced the scope of losing pre-trained features. Up to 4 epochs were used for this stage. These stages are illustrated in Figs. 2 and 3.

With this training schema, we were able to produce similar or better results for a fraction of the previous training time, using less than 20 epochs in each case. The results of training by this method are show in Table 4. A representative confusion matrix for learning validation on CIFAR-10.
Table 5  Model learning on skin data

| Model     | Average Top-1 Accuracy |
|-----------|-------------------------|
| ResNet-34 | 86.1 ± 0.6              |
| ResNet-50 | 86.9 ± 0.5              |
| ResNet-101| 89.7 ± 0.5              |
| ResNet-152| 89.4 ± 0.4              |
| DenseNet-121| 89.1 ± 0.2            |
| ResNext-50| 89.6 ± 0.8              |
| ResNext-101| 90.8 ± 1.3             |

Confusion Matrix

|      | acne | alopecia | blister | crust | erythema | leuko | macula | tumor | ulcer | wheal |
|------|------|----------|---------|-------|----------|-------|--------|-------|-------|-------|
| acne | 195  | 0        | 0       | 0     | 0        | 1     | 2      | 0     | 1     | 1     |
| alopecia | 1 | 147 | 0 | 0 | 0 | 1 | 0 | 0 | 0 | 0 |
| blister | 1 | 0 | 133 | 2 | 6 | 2 | 1 | 1 | 3 | 0 |
| crust | 1 | 1 | 2 | 138 | 0 | 0 | 2 | 1 | 4 | 0 |
| erythema | 4 | 0 | 4 | 2 | 115 | 1 | 15 | 3 | 2 | 4 |
| leuko | 1 | 0 | 1 | 2 | 142 | 3 | 0 | 0 | 0 | 0 |
| macula | 0 | 1 | 0 | 2 | 13 | 4 | 123 | 3 | 2 | 2 |
| tumor | 0 | 0 | 0 | 2 | 0 | 1 | 0 | 1 | 190 | 6 |
| ulcer | 1 | 0 | 3 | 6 | 3 | 0 | 3 | 14 | 170 | 0 |
| wheal | 0 | 0 | 0 | 0 | 1 | 0 | 0 | 0 | 0 | 149 |

Fig. 5  Confusion matrix for skin classification (Model: ResNext-101)

Fig. 6  ROC for crust, AUROC = 0.980

Fig. 7  Erythema (AUROC = 95.1%)

Fig. 8  P. Macula (AUROC = 95.7%)

3. Analysis

3.1 Effect of Sample Complexity

By systematic training of models on dermatological images, we found the validation accuracy to be converging to comparable scores across models. After obtaining best fits, we can safely hypothesize the errors emerged from the nature of data and incorrect annotations if any. To better understand the same, we focused on analyzing a few label pairs which persistently showed high rates of categorization errors (Table 6). Our investigation was based on gradient-based class activation maps (GradCAM) and Guided backpropagation (GBP) which marked the regions the model deemed important [26], [27]. In the figures from this section, the order of images (L-R) are: sample, prediction and true label.

Ulcers & Tumor

Ulcers and Tumors exhibited a high degree of errors depending on the shape of manifestation. In Fig. 9, an oval shape and minor protrusion of the tissue misled the model. Potential reasons could have been the inflammation markers surrounding lesions. Lack of adequate contrast, compounded by non-uniform illumination ranked the sample much lower on the true label. Figure 10 exhibits a case, where the model ignored the lesion’s protrusion and focused only on a section of it, incorrectly classifying it as Ulcer. In summary, Ulcers...
Table 6  Skin labels with high rates of confusion

| Labels                        | Avg. Erroneous predictions ($\mu \pm \sigma$) |
|-------------------------------|---------------------------------------------|
| Ulcer and Tumor               | 31.75 $\pm$ 6.28                           |
| P. Macula and Erythema        | 27.25 $\pm$ 3.50                           |
| Erythema and Wheal            | 19.25 $\pm$ 2.75                           |
| Crust and Ulcer               | 16.00 $\pm$ 3.46                           |

and Tumors could have a large scope of error owing to the shape and geometry.

Maculae and Erythema

Pigmented Maculae and Erythema exhibited some of the highest rates of confusion consistently across the different models we trained. For example, in Fig. 11, the redness of the lesion patch was completely ignored in favor of the surrounding skin pigmentation. These features although occupying a substantially smaller part of the total image, drew the classification in favor of the wrong class. In Fig. 12, the pigmented lesion was ignored and classified as Erythema instead based on background redness. Illumination artifacts and contrast effects are important factors in such pairs.

Ulcer and Crust

Crusts and Ulcers are chronologically related and hence pose an interesting detection challenge. We observed strong visual correlations and often the activation maps were found to be co-located. It is not surprising that models often classified incorrectly since human interpreters also exhibited a higher error rate for this pair. In the absence of patient history or other meta-information, prediction probabilities can not show strong confidence.

3.2 Robustness to Artifacts

Noise and artifacts are an inevitable part of any data centric operation. Although there are several possible means of corruption, we anticipated impulse noise and blur as two plausible candidates for user-submitted images. Impulse noise could arise by a variety of means ranging from debris on lens or imaging plate, sensor defects to uncorrected bit-flipping in storage. Movement of the imaging source or incorrect focal depth is a common observation leading to soft focus in images captured outside of controlled laboratory settings. Although human vision is capable of ignoring small defects, ML models have been known to exhibit varying levels of performance depending on data or initial conditions[28]. We chose to simulate the presence of these imperfections and calibrate the model performance.

Impulse noise is a sharp perturbation which stands out from the background as black or white pixels. The distribution of such pixels is generally random and relates to both sensor and environmental factors. Presence of such noise is known to degrade the quality of discriminative models. To simulate impulse noise, we systematically changed up to 5% of the total image pixels to black or white in the test set, via an ablation study. This set was based out of the original validation set to enable comparisons. Figure 13 is representative panel for a sample which has been corrupted with 3% impulse noise. Minimal amounts of noise did not change the images visually, with several correct predictions still (Fig. 14). The Top-1 accuracy however dropped to be-
between 64.82–73.57% in a ResNet-50 model depending on the noise quantity. The class activation maps no longer focused on the lesion entirely. There were at least nine label pairs having more than 50 classification errors among them. Figure 15 illustrates the trend in accuracy with noise addition in test set.

We performed a similarly designed ablation study involving the training data corpus. Noise was added systematically from 1–5% in the training set. The test set was kept clean. In this scenario, the accuracy observed was better. There was certain degree of resilience which was achieved by learning the model with imperfections in the training set. The results for this ablation test are shown in Fig. 16. We had an interesting observation while trying to observe the model performance in matched noise levels, i.e., the training and test set corrupted by similar quantities of impulse noise. The model prediction was found to be stable within a very narrow range of values over all the noise levels considered (81.49 ± 0.86% aggregate). These results are illustrated in Fig. 17. Details of measurements in these ablation tests are shown in Table 7. The loss in accuracy with impulse noise perturbation was seen to be higher in the case of skin images as compared to CIFAR-10. Ablation tests with similar experimental design on CIFAR-10 yielded much less fluctuation, as observed in Table 8. Skin images were seen to be quite sensitive to noise perturbation.

For simulating blurs, we used a 5×5 Gaussian kernel on images in our test set. Figure 18 is a representative panel for blur imitation induced in samples. The model was trained on regular images and tested on the modified images. In the presence of simulated blur, the accuracy reduced by approximately 12 percentage points at the minimum. The Top-1 performance was found to be 74.85% (over three trials) in the case of isotropic Gaussian blurring. A directional kernel blurring the images at 45° to the horizontal axis yielded 76.1% accuracy. The differences in outcomes between uniform and directional blurring was marginal. A confusion matrix demonstrating the model performance is shown in Fig. 19. The test accuracy difference between uniform and
directional blurring in the case of CIFAR-10 was approximately 15% (N.B: These images were blurred with $3 \times 3$ keeping in view the $32 \times 32$ size). This was significantly higher than the approximately 2% difference seen in homogeneous skin images. These results are tabulated in the final two rows of Tables 7 and 8. Based on these trials, we are led to believe that homogeneous skin images are more resilient to motion blur than noisy pixels.

3.3 Effect of Distribution Shift

In real world scenarios, medical classifiers are expected to work on diverse sets of input images. A supplied image may be different from the kind of images the model was trained on. Moreover, input images may belong to novel classes which the classifier was not designed for. In such cases, models are expected to behave predictably.

As described in Sect. 2, we created composite classes from relevant matching labels of SD-198 dataset. Approximately hundred samples were selected from each composite class for building a test set (the only exception being Wheal which had 71 samples). We trained a ResNet-50 model to optimal accuracy on our dataset, and performed inference on this test set. Poor generalization was observed, with only 32% Top-1 aggregate accuracy. Only Acne and Alopecia fared properly with 70% and 73% recall values respectively. Classifier bias was seen to favor simpler labels such as Acne and Blister dominantly over others. Figure 20 shows the confusion matrix for this test.

3.4 Response to Synthetic Data

While working on homogeneous, small datasets, we discussed data memorization as an issue with model learning. Data augmentation by generative adversarial networks (GAN) is being actively considered for this purpose [29]. The method of creating synthetic skin samples is not without its criticisms [31]. Some recent research point to the fact that generative output may not truly represent the fine visual attributes a real image has. The quality of synthetic images is very strongly related to the way the parent labels are sampled [35]. To test whether such images can truly be representative as real images, we needed to perform a
Creating synthetic data to test the robustness of models

Our first step in this process was to implement a GAN model based on label pairs which exhibited high concurrency in real clinical setting. The investigative rationale was straightforward: If synthetic samples are created from two labels in a predetermined proportion, the representation in latent space would be a linear or pseudo-linear combination of the density estimates of parent samples. When predicting a large number of such synthetic images, the ensemble accuracy should closely reflect the mixing ratio of these labels. Figure 21 is a schematic of our hypothesis.

Using approximately equal training data from Acne and Blister, which are known to exist together frequently, we designed a binary classifier. We also created several novel samples by a Progressively growing GAN (ProGAN) framework from these aforementioned labels [31], [32]. We trained the model repeatedly and calculated the posterior (prediction) probabilities for the synthetic samples each time. This ensemble performance closely followed the distribution but was not accurate enough. These results are summarized in Table 9. We have discussed the implications of this result in Sect. 4.

4. Discussion

In Sect. 2.3, we demonstrated means to perform a good model fit on homogeneous dermatological images. Despite best fits, a sufficiently large gap exists between our results and ideal, error-free performance. We ascribe it predominantly to the nature of the data itself. The visual attributes in skin diseases are not discrete by nature. Many diseases belong to wide spectrum of abnormalities which make their detection and grading difficult. Several of these disease classes are also chronologically related and different pathways could exist for lesion progression, increasing the problem complexity. Identifying intermediate transitional states between classes is a difficult challenge in machine diagnosis. Skin diseases also rarely occur in isolation. Conditions such as Erythema, Blister or Hyperpigmentation are concurrent in presentation with several other diseases. Dermatological classifiers can be biased in detecting one over the other. In the absence of valuable patient history, the information presented is only skin-deep. Unlike radiology images, skin images present very few landmarks. There are only a handful of diseases with strong visual markers. This issue is compounded by the large spectrum of skin tones in the patient pool. Lesion contrast and color are variable attributes when considering racial differences. In the absence of non-ideal images, we propose a few steps in curating a good dataset [13].

- Balancing classes in the dataset alleviates model performance. Balanced datasets exhibit a true macro-average when performing statistical analysis. In medi-
Fig. 22 Effect of having imbalanced classes in learning

Fig. 23 Optimizing field of view can immensely improve predictions

Fig. 24 Gamma balancing can achieve significant improvements in effective delineation of regions and consequently improving predictions.
ing impulse noise and simulated soft focus. Models learnt on homogeneous skin data have not yet demonstrated confidence in overcoming these small imperfections. We plan to study the effects of adversarial perturbations further to design better classification models in the future. This might also help in reducing vulnerabilities from malicious attacks in real world systems.

We are observing a dual-pronged need in building better systems: we need clinicians to gather and accurately label more information, but at the same time we require more robust behaviour in presence of noisy data. There have been several advances with respect to synthetic data, and several projects have advocated their use in computer vision fields. However, we have observed in Sect. 3.4 that the representation may not be good enough. Although the model predicted label ratios in the neighborhood of actual proportion, it was unable to guess the exact ratio in the binary representation. This error will possibly amplify in the presence of several classes. The margin of error in healthcare is slim. From our experience, computer vision techniques have been known to be useful in detecting mislabeled samples or even novel categories (Fig. 26). But a trained practitioner is required eventually to appraise this information. In light of current deficiencies, human expertise cannot be removed from the workflow.

5. Conclusion

In this paper, we demonstrated that several common skin diseases can be categorized via deep learning based classifiers from non-ideal user submitted images. We showed that given robust training, the accuracy levels become architecture agnostic. There still exists a gap between error-free detection and the peak performance achieved by such contemporary methods. This gap may not be bridged easily since it manifests from the nature of disease presentation. We also showed that the performance dipped by at least 10% in non-ideal conditions such as noise, blur and distribution shift, which are reasonable scenarios in any clinical field trial. We emphasize the role of trained practitioners in conjunction with these methods to improve the quality of dermatological services. By several scenarios presented, we conclude that unassisted autonomous diagnosis via end-to-end deep learning models is not ripe yet. They can supplement human expertise by becoming valuable physician aids for patient screening.

References

[1] Y. Liu, A. Jain, C. Eng, D.H. Way, K. Lee, P. Bui, K. Kanada, G.D. Marinho, J. Gallegos, S. Gabriele, and V. Gupta, A deep learning system for differential diagnosis of skin diseases, ArXiv preprint arXiv:1909.05382, Sept. 11, 2019.
[2] A.B. Kimball and J.S. Resneck Jr., “The US dermatology workforce: A specialty remains in shortage,” Journal of the American Academy of Dermatology, vol.59, no.5, pp.741–745, 2008.
[3] I. Dekio, E. Hanada, Y. Chinuki, T. Akaki, M. Kitani, Y. Shiraishi, S. Kaneko, M. Furumura, and E. Morita “Usefulness and economic evaluation of ADSL based live interactive teledermatology in areas with shortage of dermatologists,” International Journal of dermatology, vol.49, no.11, pp.1272–1275, 2010.
[4] H. Imaizumi, A. Watanabe, H. Hirano, M. Takemura, H. Kashiwagi, and S. Monobe, Hippocra: Doctor-to-Doctor TeleDermatology consultation service towards future AI-based Diagnosis System in Japan, IEEE International Conference on Consumer Electronics-Taiwan (ICCE-TW), pp.51–52, 2017.
[5] R.C. Lanzini, R.S. Fallen, J. Wismer, and H.C. Lima, “Impact of the number of dermatologists on dermatology biomedical research: A Canadian study,” Journal of cutaneous medicine and surgery, vol.16, no.3, pp.174–179, 2012.
[6] R.S. Stern, “Prevalence of a history of skin cancer in 2007: results of an incidence-based model,” Archives of Dermatology, vol.146, no.3, pp.279–282, 2010.
[7] T. Suneja, E.D. Smith, G.J. Chen, K.J. Zipperstein, A.B. Fleischer Jr., and S.R. Feldman, “Waiting times to see a dermatologist are perceived as too long by dermatologists: Implications for the dermatology workforce,” Archives of dermatology, vol.137, no.10, pp.1303–1307, 2001.
[8] B.A. Lowell, C.W. Froelich, D.G. Federman, and R.S. Kirsner, “Dermatology in primary care: prevalence and patient disposition,” Journal of the American Academy of Dermatology, vol.45, no.2, pp.250–255, 2001.
[9] Y. LeCun, Y. Bengio, and G. Hinton, “Deep learning,” Nature, vol.521, no.7553, pp.436–444, 2015.
[10] A. Esteva, B. Kuprel, R.A. Novoa, J. Ko, S.M. Swetter, H.M. Blau, and S. Thrun, “Dermatologist-level classification of skin cancer with deep neural networks,” Nature, vol.542, no.7639, pp.115–118, 2017.
[11] V.K. Shrivastava, N.D. Londothe, R.S. Sonawane, and J.S. Suri, “Reliable and accurate psoriasis disease classification in dermatology images using comprehensive feature space in machine learning paradigm,” Expert Systems with Applications, vol.42, no.15-16, pp.6184–6195, 2015.
[12] A.J. Park, J.M. Ko, and R.A. Swerlick, “Crowdsourcing dermatol-
ogy: DataDerm, big data analytics, and machine learning technology,” Journal of the American Academy of Dermatology, vol. 78, no. 3, pp. 644–644, 2018.

[13] S. Mishra, H. Imaizumi, and T. Yamasaki, “Interpreting Fine-Grained Dermatological Classification by Deep Learning,” Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops, pp. 2729–2737, 2019.

[14] S. Mishra, T. Yamasaki, and H. Imaizumi, “Supervised classification of Dermatological diseases by Deep learning,” arXiv preprint arXiv:1802.03752, Feb. 11, 2018.

[15] N.C.F. Codella, Q.-B. Nguyen, S. Pankanti, D.A. Gutman, B. Helba, A.C. Halpern, and J.R. Smith, “Deep learning ensembles for melanoma recognition in dermoscopy images,” IBM Journal of Research and Development, vol. 61, no. 4/5, pp. 5:1–5:15, 2017.

[16] X. Sun, J. Yang, M. Sun, and K. Wang, “A benchmark for automatic visual classification of clinical skin disease images,” Proceedings of European Conference on Computer Vision, Springer, Cham, vol. 9910, pp. 206–222, 2016.

[17] A. Krizhevsky and G. Hinton, Convolutional deep belief networks on CIFAR-10, Unpublished manuscript, Aug. 40(7):1–9, 2010.

[18] A. Paszke, S. Gross, F. Massa, A. Lerer, J. Bradbury, G. Chanan, T. Killeen, Z. Lin, N. Gimelshein, L. Antiga, and A. Desmaison, PyTorch: An imperative style, high-performance deep learning library, Advances in Neural Information Processing Systems, pp. 8024–8035, 2019.

[19] K. He, X. Zhang, S. Ren, and J. Sun, “Deep residual learning for image recognition,” Proceedings of the IEEE conference on Computer Vision and Pattern Recognition, pp. 770–778, 2016.

[20] B. Zoph, V. Vasudevan, J. Shlens, and Q.V. Le, “Learning transferable architectures for scalable image recognition,” Proceedings of the IEEE conference on Computer Vision and Pattern Recognition, pp. 8697–8710, 2018.

[21] A. Jung, ImgAug, 2017. URL: https://github.com/aleju/imgaug

[22] S. Mishra, T. Yamasaki, and H. Imaizumi, “Improving image classifiers for small datasets by learning rate adaptations,” arXiv preprint arXiv:1903.10726, March 26, 2019.

[23] L.N. Smith, “Cyclical learning rates for training neural networks,” IEEE Winter Conference on Applications of Computer Vision (WACV), pp. 464–472, 2017.

[24] I. Loshchilov and F. Hutter, SGDR: Stochastic gradient descent with warm restarts, arXiv preprint arXiv:1608.03983, Aug. 13, 2016.

[25] J. Howard and S. Ruder, Universal language model fine-tuning for text classification, arXiv preprint arXiv:1801.06146, Jan. 18, 2018.

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

[27] J.T. Springenberg, A. Dosovitskiy, T. Brox, and M. Riedmiller, Striving for simplicity: The all convolutional net, arXiv preprint arXiv:1412.6806, Dec. 21, 2014.

[28] S. Qiu, Q. Liu, S. Zhou, and C. Wu, “Review of artificial intelligence adversarial attack and defense technologies,” Applied Sciences, vol. 9, no. 5, p. 9009, 2019.

[29] I. Goodfellow, J. Pouget-Abadie, M. Mirza, B. Xu, D. Warde-Farley, S. Ozair, A. Courville, and Y. Bengio, “Generative Adversarial Networks,“ Commun. ACM, vol. 63, no. 11, pp. 139–144, 2020.

[30] B. Recht, R. Roelofs, L. Schmidt, and Y. Shankar, Do ImageNet Classifiers Generalize to ImageNet?, arXiv preprint arXiv:1902.10811, Feb. 13, 2019.

[31] A. Ghorbani, V. Natarajan, D. Coz, and Y. Liu, DermGAN: Synthetic Generation of Clinical Skin Images with pathology, arXiv preprint arXiv:1911.08716, Nov. 20, 2019.

[32] T. Karras, T. Aila, S. Laine, and J. Lehtinen, Progressive growing of GAN for improved quality, stability, and variation, International Conference on Learning Representation, April 30, 2018.

[33] P. Sprawls, Optimizing medical image contrast, detail and noise in the digital era, Medical Phys Int., 2(1), 2014.

[34] R. Sayres, A. Taly, E. Rahimy, K. Blumer, D. Coz, N. Hammel, J. Krause, A. Narayanaswamy, Z. Rastegar, D. Wu, S. Xu, S. Barb, A. Joseph, M. Shumski, J. Smith, A.B. Sood, G.S. Corrado, L. Peng, and D.R. Webster, Using a deep learning algorithm and integrated gradients explanation to assist grading for diabetic retinopathy, Ophthalmology, vol. 126, no. 4, pp. 552–564, 2019.

[35] A. Grover, J. Song, A. Kapoor, K. Tran, A. Agarwal, E.J. Horvitz, and S. Ermon, Bias correction of learned generative models using likelihood-free importance weighting, Advances in Neural Information Processing Systems, pp. 11058–11070, 2019.

Sourav Mishra received his MS from Virginia Tech in 2015 and is currently a PhD scholar in the Department of Information and Communication Engineering, Graduate School of Information Science and Technology at The University of Tokyo.

Subhajit Chaudhury received his MS from IIT Bombay in 2014 and is currently a PhD scholar in the Department of Information and Communication Engineering, Graduate School of Information Science and Technology at The University of Tokyo.

Hideaki Imaizumi received his PhD from Keio University in 2005 and is a co-founder & CEO of exMedio Inc., currently a subsidiary of MyNavi Corporation. He specializes in telemedicine and e-Health solutions.

Toshihiko Yamasaki received his B.S., M.S., and Ph.D from The University of Tokyo. He is currently an Associate Professor at Department of Information and Communication Engineering, Graduate School of Information Science and Technology, The University of Tokyo. He was a JSPS Fellow for Research Abroad and a visiting scientist at Cornell University. His current research interests include attractiveness computing based on multimedia big data analysis and fundamental problems in pattern recognition and machine learning. He is a member of IEEE, ACM, AAAI, IEICE, IPSJ and ITE.